Office of Naval Research
Contract N00014-82-K-0727

covering the period
1 July 1984 - 30 June 1985

ANNUAL PROGRESS REPORT

Speech Recognition: Acoustic, Phonetic and Lexical

Office of Naval Research
Contract N00014-82-K-0727

covering the period
1 July 1984 - 30 June 1985

Submitted by:
Victor W. Zue

August 5, 1985

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Research Laboratory of Electronics
Cambridge, Massachusetts 02139

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release; distribution unlimited

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)
ARPA Order No. 4585

6a. NAME OF PERFORMING ORGANIZATION
Research Laboratory of Electronics
Massachusetts Institute of Technology

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION
Advanced Research Projects Agency

6c. ADDRESS *(City, State and ZIP Code)*
77 Massachusetts Avenue
Cambridge, MA 02139

7b. ADDRESS *(City, State and ZIP Code)*
1400 Wilson Boulevard
Arlington, Virginia 22217

8a. NAME OF FUNDING/SPONSORING ORGANIZATION
Office of Naval Research
Mathematical and Physical Sciences Res. Progr.

8b. OFFICE SYMBOL *(If applicable)*

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
N00014-82-K-0727

8c. ADDRESS *(City, State and ZIP Code)*
800 North Quincy Street
Arlington, Virginia 22217

10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
	NR 049-542		

11. TITLE *(Include Security Classification)*
Speech Recognition: Acoustic, Phonetic and Lexical Knowledge

12. PERSONAL AUTHOR(S)
by Victor W. Zue

13a. TYPE OF REPORT
Annual Report

13b. TIME COVERED
FROM 7/1/84 TO 6/30/85

14. DATE OF REPORT *(Yr., Mo., Day)*
5 August 1985

15. PAGE COUNT
428

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB. GR.

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

Work by Victor W. Zue and his collaborators is summarized here.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. □ DTIC USERS □

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Kyra M. Hall
RLE Contract Reports

22b. TELEPHONE NUMBER *(Include Area Code)*
(617)253-2569

22c. OFFICE SYMBOL

EDITION OF 1 JAN 73 IS OBSOLETE.

SECURITY CLASSIFICATION OF THIS PAGE

RESEARCH LABORATORY OF ELECTRONICS, 36-549

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

CAMBRIDGE, MASSACHUSETTS 02139

August 5, 1985

Director
Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22200

Attention: Program Management

This letter is the Annual Progress Report for our research program supported under DARPA-ONR Contract N00014-82-K-0727. → During this reporting period we

During the period of July 1, 1984 to June 30, 1985, we have continued to make progress on the acquisition of acoustic-phonetic and lexical knowledge. Specifically:

(1) • We ~~have~~ completed ~~the~~ development of a continuous digit recognition system. The system was constructed to investigate the ~~utilization~~ use of acoustic-phonetic knowledge in a speech recognition system. The significant achievements of this study include the development of a "soft-failure" procedure for lexical access and the discovery of a set of acoustic-phonetic features for verification.

(2) • We ~~have~~ completed a study of the constraints that lexical stress imposes on word recognition. We found that lexical stress information alone can, on the average, reduce the number of word candidates from a large dictionary by more than 80 percent. In conjunction with this study, we successfully developed a system that automatically determines the stress pattern of a word from the acoustic signal.

(3) • We ~~have~~ performed an acoustic study on the characteristics of nasal consonants and nasalized vowels. We have also developed recognition algorithms for nasal murmurs and nasalized vowels in continuous speech.

→ to p. ii

i

(4) • We ~~have~~ finished the preliminary development of a system tnat aligns a speech waveform with the corresponding phonetic transcription.

We are including with this report copies of all publications, in the form of theses and papers presented at various conferences, written during this contracting period.

Sincerely yours,

Victor W. Zue
Principal Investigator

Enc.

VWZ/kk

Lexical Stress and its Application in Large Vocabulary Speech Recognition

by

Ann Marie Aull
B.S.E.E., Purdue University (1982)

Submitted in partial fulfillment of the requirements for the degrees of

Master of Science and Electrical Engineer

at the

Massachusetts Institute of Technology

August 1984

Signature of Author ..... Ann Marie Aull ....
Department of Electrical Engineering and Computer Science
29 August 1984

Certified by ...... V. W. Zue ....
Thesis Supervisor

Accepted by ....
Chairman, Departmental Committee

Lexical Stress and its Application in Large Vocabulary Speech Recognition

by

Ann Marie Aull

Submitted to the
Department of Electrical Engineering and Computer Science
on 29 August 1984 in partial fulfillment of the requirements
for the Master of Science and Electrical Engineer Degrees

Abstract

This thesis addresses the issue of lexical stress determination from the acoustic signal. The motivation for this work stems from the fact that stressed syllables provide islands of acoustic-phonetic reliability, which is valuable for most phonetically-based recognition systems. A further motivation for stress determination stems from the constraining power of stress information in a particular recognition scheme based on lexical access. Past lexical studies suggest that the lexical constraint provided by segments around stressed syllables is stronger than that around unstressed syllables. To investigate the application of stress information in a specific recognition scheme, the lexical constraints provided by stress information are examined. In particular, a large vocabulary isolated word system that performs lexical access based on partial phonetic information provides a potential application for prosodic information. To examine the lexical constraints provided by stress, the polysyllabic words in the Merriam Pocket Dictionary are mapped into their corresponding stress patterns. The results indicate that, from stress information, the largest class size constitutes 28% of the lexicon. The expected value of the class size corresponds to 15% of the lexicon illustrating the constraining power of the stress information.

In order to locate regions of acoustic reliability and to exploit the lexical constraints, we develop a system that determines the stress pattern from the acoustic signal of isolated words and performs subsequent lexical access. The system initially segments the speech signal into broad phonetic classes. From these segments, sonorant regions of the syllables are determined. Next, known acoustic correlates of stress, such as duration, energy, and fundamental frequency, are extracted for each syllable. The stress pattern is established through a relative comparison of the syllable feature vectors. Finally, lexical access based on the derived stress pattern provides a list of word candidates. Phonological rules are incorporated to account for variations in the number of syllables from the lexical base forms. The system is evaluated on a database of over 1600 isolated words, spoken by 11 speakers, with varying degrees of difficulty for deriving the syllable units. The system performance in establishing the correct stress pattern is 87% accurate. Of the 13% error, 10% is due to confusion between unstressed and reduced syllables or to not finding the correct number of syllables. Only 3% is due to labeling the stressed syllables as unstressed.

Thesis Advisor: Victor Zue
Title: Assistant Professor of Electrical Engineering and Computer Science

Acknowledgments

I would like to thank several people for their interest and contributions throughout my thesis work. First, my sincerest thanks to my thesis advisor Victor Zue for providing for me the opportunity and the environment to pursue this research. He has been a constant source of support, guidance, and, most importantly, friendship. Working with Victor is a most rewarding as well as challenging experience. I also extend my thanks to Hong Leung who provided many preprocessing programs in addition to his companionship. Thanks to those in 36-547 for not being in their office this summer as it was a quiet place to do a lot of writing. I would like to thank Stephanie Seneff who wrote the pitch detection algorithm incorporated in my work. I would also like to thank Lori Lamel, Jim Glass, and Dan Huttenlocher for access to their databases. Finally, I extend my gratitude to the members of the Speech Communication group for their input.

My deepest appreciation is for my family. My parents, Wilma and Roger, have always been my inspiration. Their love, encouragement, and belief in the value of hard work have provided me with the ability to do well in all my endeavors. My brothers have also given me love and guidance and provided success stories that I could try to emulate in my work. Finally, my thanks to Eddie, whose continual love and support gave me the strength to do well, who could renew my confidence and uplift my spirits.

To my parents, Wilma and Roger

Biographical Note

Ann Marie Aull was born in Indianapolis, Indiana, on August 7, 1960. After graduating from North Central High School as valedictorian, she attended Purdue University in West Lafayette, Indiana, where she received a Bachelor of Science in Electrical Engineering with highest honors. She began her graduate education in September, 1982, at the Massachusetts Institute of Technology where she is expected to complete the Master of Science degree in August, 1984.

As an undergraduate, she participated in the cooperative plan of education for two years at General Motors, Detroit Diesel Allison Division. Two summers were spent at Bell Laboratories in the Home Communication Lab doing C programming for various projects.

She hopes to work in the industry in the areas of speech recognition and digital speech processing. After some work experience, she plans to return to school and begin work on a Ph.D.

Table of Contents

Chapter One

Introduction

1.1 Review of Speech Recognition

The field of speech recognition has seen considerable advancements in the last years as we continually strive for a graceful interaction between man and machine. The ARPA SUR project in 1971 provided the direction and funding for a large-scale effort in continuous speech recognition systems. Two of the most successful systems to come out of the five year project, namely HEARSAY II and HARPY, incorporated a representation of all levels of speech and linguistic knowledge, i.e. acoustic-phonetic, phonological, lexical, syntactic, and semantic [13]. The success of the ARPA project opened to the speech community a new perspective and gave directions to further pursue in continuous speech recognition as well as in isolated word recognition.

The implication of the ARPA project for further speech research concerns the phonetic characterization of the speech signal. The SUR systems emphasized top down processing such as the use of syntactic and semantic information to compensate for a poor acoustic and phonetic characterization of the speech signal. It was believed that the speech signal was inherently too variable or noisy to extract robust low-level information. However, more recent spectrogram reading experiments [4] have renewed the scientific community's interest in phonetic recognition. These experiments have shown that an appreciable amount of acoustic-phonetic information may be derived from the speech signal. These studies hoped to qualify the low level information carried in the speech signal and investigated how speech scientists can take advantage of that knowledge to build better

recognition systems. From these experiments the role of knowledge such as acoustic-phonetics, syntax, and semantics, is more clearly understood. At the same time, the studies examine the amount of invariant information that could make spectrogram reading possible and improve the acoustic knowledge sources of speech recognition systems. A recent investigation by Cole et al [4] show overall results that an expert spectrogram reader can achieve 85% accuracy on phonetic recognition with only the use of acoustic and phonetic information (as compared with 40% in the SUR systems [13]). In the experiment, higher level information, such as syntax and semantics, played a role not in segment identification but in word and sentence hypothesizations from the phonemic transcription.

Spectrogram reading experiments shed light on the knowledge representation that a human expert uses which may be incorporated in the construction of continuous speech recognition systems. However, spectrograms do not readily illuminate suprasegmental information such as timing, intonation, or stress. Stress is a basic prosodic feature which, as discussed by Lea [15], is an important component for speech recognition to:

- distinguish word pairs such as *CONtract* and *conTRACT*;
- provide pointers to islands of phonetic reliability;
- locate areas of emphasis and semantic contrast, and disambiguate syntactically ambiguous sentences;
- determine intonation;
- derive conditions for the application of phonological rules.

Sentential stress remains elusive due to complicated effects such as declination, intonation, and phonetic distortion. At the sentence level, one must also deal with varying levels of stress such as the overall sentential stress, contrastive stress, phrase level stress, and lexical stress. Lexical stress presents a more focused subproblem that serves as a basis for further prosodic study. In this thesis, the role of lexical stress in speech recognition will be explored from two perspectives. First, the lexical constraints of stress information are further investigated. Second, a stress recognition system is implemented.

1.2 Functional Utility of Lexical Stress Information

The importance of lexical stress for large vocabulary speech recognition is motivated by a series of lexical studies by Huttenlocher and Zue [9]. They investigate the sources of constraint for a lexicon of 20,000 words, first from segmental information alone and then from the combination of segmental with prosodic information. In their studies, each word in the lexicon is associated with only one baseform pronunciation. The phones of each word in the lexicon are classified into broad phonetic classes: stop, weak fricative, liquids and glides, strong fricatives, nasals, and vowel. A *cohort* is defined to be the equivalence class of words that match a particular sequence of symbols. If the cohort size is one, then that pattern uniquely specifies the word in the lexicon. An expected equivalence class size, as opposed to a mean, takes into account the nonuniform distribution of words across classes. The expected class size, E(class size), is computed as

$$E(class\ size) = \sum_{i=1}^{n} C_i \ P_i$$

where C_i = size of class i

P_i = probability of choosing a word from class i

n = number of classes

The expected class sizes are calculated assuming first an equal frequency weighting of the words in the lexicon. Second, the expected class size is computed incorporating the frequency of usage in the English language as determined from the Brown corpus. The results shown in Table 1-1 indicate that segmental information provides strong lexical constraints as the expected class size constitutes only .2% and .1% of the lexicon with and without frequency weighting, respectively. If a phonetic sequence were chosen at random, without frequency weighting, one would expect a unique correspondence almost one third of the time, and a total of 223 words corresponding to the maximum class.

The above results are interesting in that a broad class representation is both practical and can provide such powerful constraints. It is more practical to classify the speech signal into these categories corresponding to *manner of articulation* as opposed to segmentation at the detailed phonetic level. Sound units that correspond to a particular manner of articulation possess characteristics that are acoustically robust and relatively invariant with respect to other manner classes and across speakers. For example, the phonemes that are classified as stops, /p,t,k,b,d,g/, consist of a closure or silence followed by a release and frication. These traits remain fairly robust across speakers and allophones of that phoneme.

	No Weight	Frequency Weighted
Expected Class Size	21	34
Median Class Size	4	25
Maximum Class Size	223	223
% Uniquely Specified	32%	6%

Table 1-1: Summary of Results of Broad Phonetic Classification from Huttenlocher and Zue

The advantage of the broad classes lies in its ability to maintain the phonotactic constraints and account for allophonic variation. However, one must also deal with phonetic variability such as segment deletion, insertion, or transformation. For now, multiple lexical entries to incorporate various pronunciations are impractical due to storage and search limitations. Also, this representation does not take advantage of the fact that some segments are extremely variable while others remain relatively intact. Most of the variability occurs in unstressed syllables while stressed syllables remain intact. Huttenlocher and Zue [9] further examine this phenomena. The phones are again mapped into one of the six broad classes but stress information is also included. As shown in Table 1-2, segmental plus prosodic information provides greater constraint than segmental information alone as demonstrated by the reduced class sizes. In an additional experiment, two types of classifications are performed. First, only the phones in the *stressed* syllables are mapped into phonetic classes.

Unstressed phones are simply marked as a place holder. Second, the lexicon is mapped with *unstressed* syllables classified leaving stressed phones only as place holders. The classification and its results are shown in Table 1-3. These results suggest that stressed syllables, which tend to be less phonetically variant, provide the majority of constraint for a large lexicon. Accounting for phonetic variability becomes imperative in multisyllable words, such as *international*, that have various pronunciations as seen in the spectrograms of Figure 1-1. Yet the stressed syllables remain robust acoustically. Thus, the results of these studies suggest that the combination of segmental and prosodic information provides more constraint than segmental alone, where the majority of constraint is due to information around stressed syllables.

	Segmental Only	Segmental + Stress	Segmental Frequency Weighted	Segmental + Stress Frequency Weighted
Expected Class Size	21	16	34	20
Median Class Size	4	2	25	8
Maximum Class Size	223	209	223	209
% Lexicon Unique	32%	30%	6%	27%

Table 1-2: Summary of Results of Segmental and Prosodic Classification from Huttenlocher and Zue

Representation of a large lexicon in terms of broad phonetic classes and stress information has the following advantages.

	No Deletion	Delete Unstressed	Delete Stressed
Expected Class Size	21	40	2013
Maximum Class Size	223	261	3703
Median Class Size	4	22	1725

Table 1-3: Summary of Results of Classification of Broad Phonetic Information in Stressed and Unstressed Positions from Huttenlocher and Zue

- Broad classes are accustically robust and invariant with respect to other classes and across speakers.

- Some of the allophonic variations, such as the /t/ in *Tom, tree* and *stay*, are retained in the broad classification.

- Phonotactic constraints are maintained even with the broad classes. For example, a weak fricative can not precede a nasal in a word-initial position.

- Phonetic variability such as deletion usually is restricted to unstressed syllables leaving stressed syllables highly invariant.

- The lexical search and storage space is greatly reduced.

- The representation may be extendible to continuous-speech, speaker-independent speech recognition. Such methods have motivated the implementation of a similar representation in a speaker-independent, connected-digit recognizer by Chen [2].

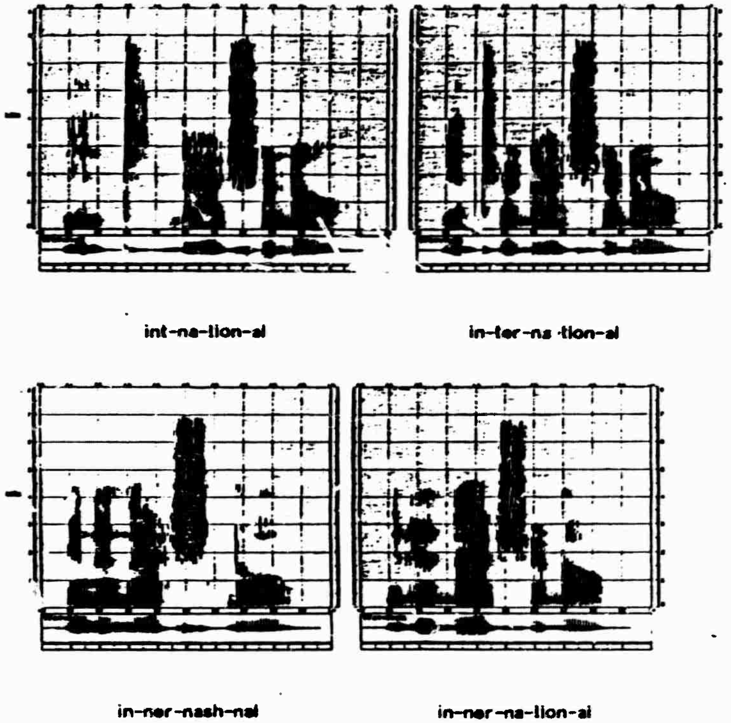


Figure 1-1: Spectrograms illustrating several possible pronunciations for the word *international*

There is strong evidence that prosodic information serves two major purposes for speech recognition. First, lexical stress provides a significant amount of lexical constraint that reduces the search space for large vocabulary lexical access. Second, stress determines where the speech signal is rich in acoustic information. Together, these benefits render stress an important component in advanced speech recognition systems.

1.3 Past Studies of Stress

Stress is important for continuous speech recognition as it embodies higher level information that affects the semantic, syntactic, phonological, and phonetic content of the speech. However, the focus of this thesis will be on lexical stress in order to reduce the complications imposed by higher level information. By reducing higher level effects manifested in continuous speech, it is hoped that the acoustic correlates reflect stress alone and not rhythm, intonation, or other suprasegmentals. The majority of work discussed in this section is related to lexical stress in isolated words and does not survey the more complicated problem of prosodic events in continuous speech. This section covers the past studies of the acoustic correlates of lexical stress, the perception of lexical stress, and the recognition of stress.

1.3.1 The Acoustic Correlates of Lexical Stress

The acoustic manifestation of stress as discussed by Lehiste [16] is judged in terms of effort. Rather than a single parameter corresponding to the production of stress, the acoustic correlates of respiratory effort are incorporated in the concepts of energy, duration, fundamental frequency, and intensity. An increase in effort can produce a higher rate of vocal fold vibration and greater subglottal pressure which yields more power along the sound wave. Intensity is thus increased as it is a measure of the power transmitted along the sound wave through an area perpendicular to the direction of propagation. Lehiste claims that duration is not directly related to effort but remains a language-determined phenomenon. It could

be argued that duration is actually a covariable of respiratory effort in that in order to produce a certain quality the effort must be sustained for a minimum period of time. A suitable analogy would be the inherent lengthening of a tense vowel over a lax vowel.

The influence of intensity, fundamental frequency, and duration in the perception of stress have been studied by Fry [7] [8], Bolinger [1], and Morton and Jassem [20]. The study by Fry [7] investigated the effect of the acoustic cues of duration and intensity on the perception of word stress. The corpus of words consisted of noun/verb pairs such as *CONtract/conTRACT*. Fry varied the duration and intensity of the vowel in the first syllable with respect to the vowel of the second syllable (V1/V2). Listeners were then asked to mark where they perceived the stress in the synthesized data. The range of duration and intensity values reflected actual measurements attained from recordings of the word corpus. The recorded data also showed that the majority of change in these parameters with a shift of stress is reflected in the vowel regions as opposed to the surrounding consonants. The results showed that both duration and intensity are acoustic cues for the judgment of stress. The results also indicated that the duration ratio has a stronger influence on the perception of stress than intensity as depicted in Figure 1-2.

In the 1958 study, Fry analyzed intensity, duration, and fundamental frequency as acoustic properties for the perception of stress. The variation of the fundamental frequency was controlled by a device on which hand painted spectrograms were used as the input to a synthesizer. The duration ratios of the previous experiment were combined with incremental changes in fundamental frequency, ranging from 5 to 90 Hz, keeping the intensity ratio constant. The results indicated that an increase in fundamental frequency was more likely [illegible] perceived as stressed independent of the magnitude of the change.

Bolinger's investigation [1] of the properties of stress led him to the conclusion

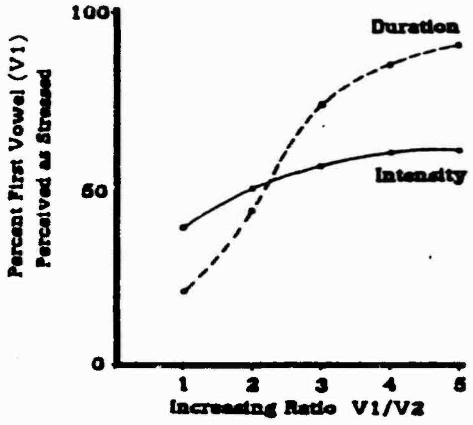


Figure 1-2: Percentage of the first vowel, V1, to be stressed for all test words as a function of vowel duration ratio and vowel intensity ratio as studied by Fry

that pitch was the prominent cue. Studying both natural and synthetic speech, he claimed that duration is a suitable covariable of pitch but intensity does not play a major role. However, the inconsistency of Bolinger's results with respect to intensity might be attributed to the intrinsic intensity of phonetic segments which was ignored by Bolinger. In an attempt to eliminate such an effect, Morton and Jassem [20] used synthetic nonsense words such as *sisi* or *sasa* to maintain uniform vowel quality in both syllables. Intensity, duration, and fundamental frequency were systematically varied similar to Fry's experiments. The listeners perceived stress on syllables that were more intense and longer. Consistent with Fry's findings, a rise in fundamental frequency was perceived as stressed without regard for the absolute magnitude.

1.3.2 Perceptual Studies of Lexical Stress

There is perceptual evidence that stress provides an anchor for word understanding and verification. In a study by Cole and Jakimik [5], a story was read to the subjects with mispronounced words inserted. The subjects noted when they detected a mispronunciation. The results showed that mispronunciations were detected more often in stressed than unstressed syllables independent of the position of the stressed syllable in the word. Similarly, a study by Cutler and Foss [6] investigated the effect of stress on reaction time to word-initial phoneme targets on content and function words in sentence contexts. The stress level of the words was varied as well as the overall sentence stress pattern. They found that the reaction times were shorter for the stressed words independent of the their syntactic role or the *normalcy* of the sentence stress pattern. Thus, the perceptual studies lend support to the notion that stress plays a major role in sentence and word comprehension.

1.3.3 Recognition of Lexical Stress

The acoustic correlates of stress have received much attention in the area of speech synthesis in an attempt to create more natural sounding speech. In this task, stress information is readily incorporated, as the phonetic environment is known *a priori*. The problem of normalizing for inherent duration and intensity of segments such as vowels is thus eliminated. The recognition of lexical stress, however, remains elusive as phonetic information can not be assumed. The acoustic cues for stress are often ambiguous and certainly influenced by the phonetic environment. A *measure* of stress is an abstraction that may not exhibit robust physical properties.

One attempt to build a stress recognition system was made by Lieberman [18]. In this experiment, two syllable noun/verb word pairs such as *OBject/obJECT* were recorded. Listeners judged the stress location of the corpus. The relation of the perceived stress to the parameters of fundamental frequency, peak envelope amplitude, duration, and the integral of the amplitude were investigated. The stressed syllable had a higher maximum fundamental frequency than the unstressed of the same token in 90% of the cases, a higher peak amplitude in 87%, a longer duration in 66%, and a higher integral of amplitude in 92%. The stressed syllable compared to its unstressed counterpart in the other word of the pair had a higher fundamental frequency in 72% of the cases, higher peak amplitude in 90%, and a longer duration in 70%. In no case did the stressed syllable have both a lower amplitude and lower maximum fundamental than the unstressed. In addition, in nearly 100% of the data, the stressed syllable had either a greater integral of amplitude or a longer duration than the unstressed syllable of the same word. With this information a program for automatic recognition of stress for a two syllable word input was constructed as shown in Figure 1-3. Fundamental frequency and envelope amplitude were emphasized as cues for stress. Judgments by this system agreed with the perceptual judgments 99% of the time. The relatively poor performance of duration as an acoustic cue can be partially attributed to segmental effects. As reported by Klatt [11], a vowel in a syllable that is in word and phrase final position may be lengthened an average of 30%. Many of the words in the corpus fell at the end of the carrier sentence. In addition, isolated words may be subject to additional lengthening. The corpus of words were in a sentence structure yet the subjects were asked to read the sentences silently except for the underlined data word. Thus, duration as a cue for stress may be diminished without compensating for prepausal lengthening.

The reliability of fundamental frequency as an acoustic cue for stress again depends on syllable position. Words spoken in isolation often have a rise in fundamental frequency on the final syllable which could influence the decision of an automated scheme. Words embedded in a carrier phrase would more than likely show a more natural, gradual decrease or declination in fundamental frequency.

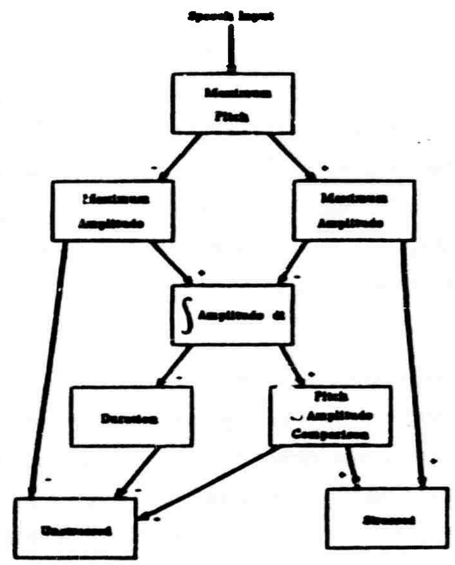


Figure 1-3: Automatic Recognition of Stress in Two Syllable word input as designed by Lieberman

An important issue related to Lieberman's experiment is the degree of stress. Bisyllabic words have a relatively clear stress pattern. Polysyllabic words such as *carbohydrate* may have levels of primary and secondary stress as well as unstressed and reduced. As the difficulty in perception of stress increases, the ambiguity of the acoustic correlates likewise increases. Even though Klatt [11] reports only a 5% duration difference between primary and secondary stressed vowels, the difference between unstressed and secondary stress remains ambiguous both acoustically and perceptually.

In developing an algorithm, the proper combination of parameters remains as important as the determination of robust parameters. The abstract nature of stress may lend itself to a heuristic approach to the construction of an automatic stress recognizer. Lieberman carefully examined the measured data and from the information developed an effective heuristic for determining lexical stress in a two syllable word. Such a development technique may be extendible to polysyllabic words even if the specific algorithm is not. As a final note on Lieberman's study, part of the success of his system might be attributed to the fact that the training data was also used for the system evaluation. In this situation, it is hard to avoid tweeking the system to the data. Thus, it is the intent of this thesis to derive the stress algorithm from one set of data and to then evaluate on a completely new database in order to establish an objective measure of the system performance.

1.3.4 Summary

It is clear from past work that stress has acoustic correlates of duration, energy, and fundamental frequency. In addition, there is perceptual and acoustic evidence that stress provides pointers to regions of robust information. However, for purposes of recognition, there is ambiguity as to how to combine the acoustic correlates and derive a stress decision without detailed phonetic information. Part of the strength in Lieberman's system lies in the fact that the stress decision is relative to only the other syllables in the word. The importance of this concept is that different people speak with different emphasis, rates, and articulation. The relative decision acts as a normalization for these effects which cannot be accomplished by some global normalization. In addition to speaker variability, there is inherent phonetic variability. For example, a stressed lax vowel in one word

may be shorter in duration than an unstressed tense vowel in another word. A comparison of these two vowels would not be meaningful. A relative decision for stress again helps to normalize for intra-speaker variability as well as inter-speaker variability.

Lieberman focused on deriving stress for two syllable words and a limited number of speakers. The success of stress information for future speech recognition systems depends on the ability to achieve complete speaker independence for any multisyllable word.

1.4 Overview of the Thesis

Current state of the art technology in speech recognition systems relies heavily on pattern recognition based on stored word or phoneme templates. However, this technology may not be extendible to large vocabulary systems or speaker independence. In particular, the incorporation of stress into the template scheme remains elusive as much of the durational information important to stress may be lost in the dynamic time warping algorithm common to these systems. Therefore, this thesis intends to take a phonetic approach to lexical stress. The determination of the stress pattern for a given isolated word is the primary focus of the thesis. More specifically, a front-end processor establishes sonorant regions of the syllables. The parameters best illuminating lexical stress are extracted from these regions and input to a stress algorithm.

A prosodic identification component could be integrated into large vocabulary isolated word recognition systems. Such a system has been proposed by Huttenlocher based on segmental information. The input speech signal is characterized into broad phonetic classes. Word candidates are proposed by means of lexical access in which the output of the broad segmentation process is mapped against the broad phonetic representation of the words in a 20000 word lexicon. This system, however, embodying no prosodic information, relies on segmental information alone. Lexical studies have established the contribution of prosodic information to such a system and motivates the development of a prosodic component. The implementation of a stress recognition system alone does not imply a means of speech recognition; rather, it is the ultimate combination of prosodic and segmental information that provides an acoustically robust and well constrained framework for speech recognition.

In Chapter 2, the well-studied acoustic correlates of stress, namely duration, energy, and fundamental frequency, are examined from a recognition standpoint similar to some of the past work discussed previously. Also, we examine the lexical constraints provided by stress information for a particular recognition task. In contrast to the studies by Huttenlocher and Zue in which segmental information is combined with prosodics, we probe the constraining power of stress information alone.

Chapter 3 concerns the implementation of a stress recognition scheme. An outline of the system is presented and each of the components is subsequently discussed in detail. A front-end processor establishes syllable regions. The stress algorithm extracts features and makes a decision on the stress pattern for the word. Finally, the lexical access component presents a list of word candidates for the derived stress pattern based on a large lexicon that has been expanded by a set of phonological rules. In this representation, one word may have several stress patterns associated with it. The lexical component is not a proposal for an appropriate construction of a recognition system, but merely demonstrates some of the lexical properties discussed in Chapter 2.

In Chapter 4, we show the results of the system evaluation and performance. The evaluation is broken down into the separate components of the system as well as by

degree of difficulty of the evaluation data. Thus the performance of the stress recognition system is thoroughly examined.

Chapter 5 presents a summary of the thesis. In addition, suggestions for further research are discussed.

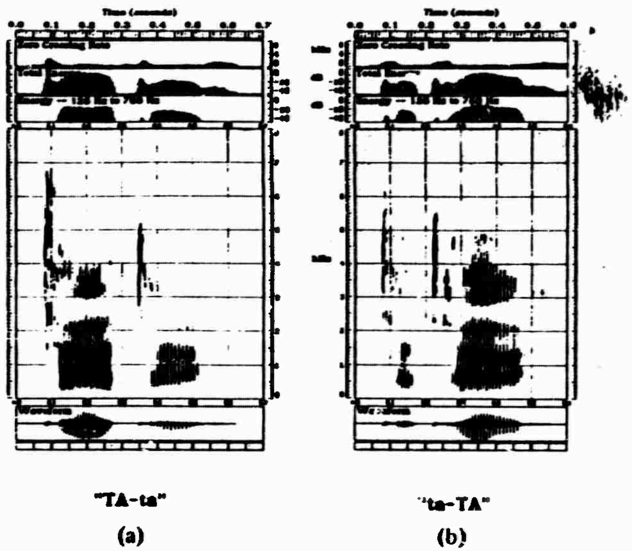
Chapter Two

Acoustic and Lexical Properties of Stress

In the previous chapter, we discussed the importance of prosodic information for speech recognition. Limiting ourselves to the subproblem of isolated words as opposed to continuous speech, lexical stress becomes the primary focus of attention, eliminating the additional complexities of sentential stress. There is evidence from the previous work discussed that lexical stress may serve at least two purposes for speech recognition: in general, stressed regions are more acoustically robust than unstressed; in a particular recognition scheme, stress provides a great deal of lexical constraint.

As seen in Chapter 1, there may be physical correlates of stress that can be extracted from the speech signal to identify regions of stress and thus regions of acoustic reliability. In other words, one can be more confident in identifying phonetic events in a stressed environment than an unstressed. This information would be beneficial, in general, to most phonetically based recognition systems. As an example, the nonsense word *ta-ta* is spoken first with the stress on the first syllable and then on the second. In the wide-band spectrograms shown in Figure 2-1, both syllables have the same phonetic composition, yet the stressed syllable in both cases displays more information such as formant locations and burst characteristics.

Stress information may play an important role in the particular application of a large vocabulary, isolated word recognition system. For example, the system proposed by Huttenlocher [10] segments the speech signal into broad phonetic classes. Word candidates are proposed by mapping the phonetic string into a 20000 word lexicon where each word in the lexicon has been characterized into its broad phonetic sequence of events. In such a system where lexical access is used to present word candidates based on partial phonetic information, lexical stress may constrain the task in several ways. First, knowing the location of stress immediately eliminates words that may have the same broad phonetic representation but different stress patterns. For example, the words *campus* and *compose* both consist of the broad phonetic sequence [Stop Vowel Nasal Stop Vowel Strong-Fricative]; however, the



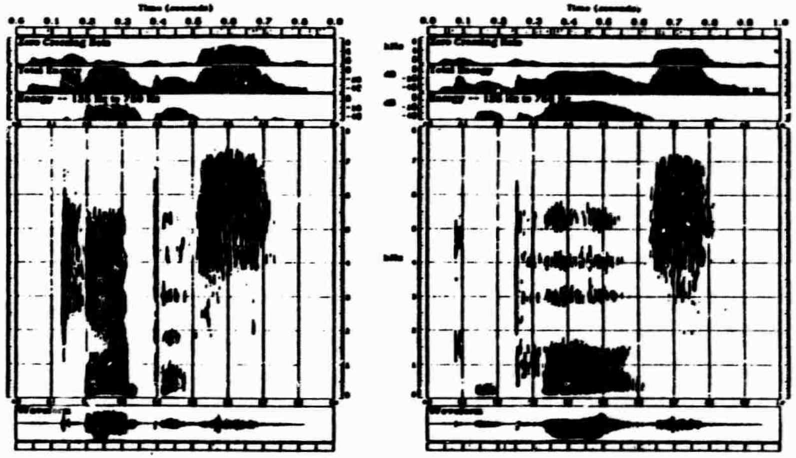
"TA-ta"
(a)

"ta-TA"
(b)

Figure 2-1: Wide-band spectrograms of the nonsense word *ta-ta* spoken with (a) stress on the first syllable and (b) stress on the second syllable

lexical stress remains a distinguishing feature as the stress falls on the first syllable in *campus* and on the second in *compose*. See Figure 2-2 for wide-band spectrograms of these words. Second, the lexical studies of the previous chapter show that phonetic information around stressed syllables provides more constraint than that around unstressed. The derived phonetic information around the stressed syllables could be weighted more heavily than that derived for the more variable unstressed syllables, eliminating some of the front-end processing errors that could be propagated in such a recognition system.

Past studies have indicated that the acoustic correlates of stress depend not on the detailed phonetic form, but on suprasegmental properties. Thus, the acoustic measurements can be made once the vowel regions have been located. From a practical standpoint, however, it is often difficult to delineate the vowel from adjacent sonorants (as in the word *conceal*). Thus, one must wrestle with the practical problem of how to obtain a robust syllable segmentation while mindful of extracting meaningful measurements. It is the intent of this chapter to examine some of the acoustic correlates of stress that can be extracted directly from the speech signal from two perspectives. First, we assume that the vowel boundaries are known. Second, and more practically, we assume our information is limited and only the boundaries of the sonorant portions of the syllables are known. This comparison helps us to determine the potential of building a practical recognition system for stress determination. In addition, the lexical constraints provided by stress are investigated for the task of large vocabulary, isolated word recognition. These studies are an attempt to confirm our belief that a stress recognition component is both feasible and meaningful for speech recognition.



"campus"
(a)

"compose"
(b)

Figure 2-2: Wide-band spectrograms of two words consisting of the same broad phonetic sequence but different positions of stress as (a) *campus* has stress on the first syllable and (b) *compose* has stress on the second syllable

2.1 Measurement and Evaluation of Acoustic Correlates of Stress

The database for initial acoustic measurements consists of a total of approximately 350 polysyllabic words spoken by three female and four male speakers. The fifty word corpus is listed in Appendix A. The majority of words are two, three, and four syllable with varying stress patterns and phonetic contexts. The distribution of the corpus by number of syllables is shown in Figure 2-3. The words in the database are recorded, digitized at 16 kHz, and time aligned with their phonetic transcriptions using the SPIRE facility [23].

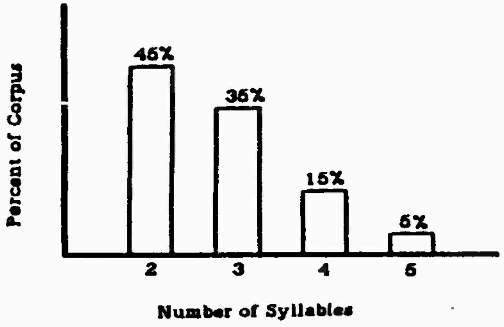


Figure 2-3: Distribution of corpus by number of syllables

In order to analyze the data for correlates of lexical stress, the *correct* stress for each word must first be established. An objective means of determining the appropriate stress is through listening tests. Five subjects were individually asked to listen to a recording of the words where each word was repeated three times. They were asked to mark which syllable received the greatest stress for each word, or to mark the word as ambiguous if no clear emphasis could be perceived. From the

listeners' results, the stress for each word was assigned by the decision of the majority of listeners. When a majority agreement could not be established for a word, it was removed from the database. As a result, a total of fifty words was deleted from the database. By associating each word in the corpus with a perceived stress pattern, we are more confident that the acoustic correlates to be studied are directly associated with the perceived stress for each word.

There is evidence that lexical stress influences the acoustic properties of a syllable [12]. For example, vowels are longer in stressed contexts than unstressed. In addition, prestressed consonants have a longer duration than their unstressed counterparts. However, vowels are probably the most greatly influenced by stress due to the nature of vowel articulation. The open vocal tract configuration for a vowel is easier to manipulate than that of a constricted one such as in the production of stops or fricatives. Another reason to concentrate on the effect of stress on vowels is that every syllable has a vowel (or syllabic) segment whereas the number of surrounding consonants is highly variable from syllable to syllable. Thus, if one were to compare syllables for stress determination, focusing on vowel segments provides some normalization across syllables.

Ideally, the vowel regions would best illuminate the influence of stress. This acoustic study deals with transcribed data where vowel and syllable boundaries are marked and aligned with the speech waveform. As discussed previously, a crude initial segmentation is more robust with respect to phonetic and speaker variability. Labeling the speech into classes such as sonorant or fricative is less difficult and more practical than distinguishing all of the sonorants, i.e., vowels, nasals, liquids, and glides. Often, it may be extremely difficult to extract vowels from such surrounding sonorants. This study, then, also investigates the influence of stress on the *sonorant* regions of the syllables in comparison to the vowel regions.

Each word in the database is associated with its time-aligned phonetic transcription which includes syllable boundaries. Therefore, the boundaries for either the vowel region for each syllable or the sonorant region for each syllable are known. In this acoustic study, two sets of parameter measurements are evaluated. The first is on the vowel regions of the data and the second is on the sonorant regions of the syllables.

The well studied correlates of duration, energy, and fundamental frequency are computed on each utterance in the corpus and extracted from the vowel and sonorant regions. These features[1], are interrogated as cues for lexical stress. Duration measurements are made on the vowel regions and on the sonorant regions of the syllables based on the time boundaries derived from the hand labeled phonetic transcription. Since the speech is hand transcribed by examination of the sampled waveform, the inherent accuracy of the boundaries is ideally limited by the sampling rate. The speech is windowed with a 6.7 msec Hamming window (300 Hz bandwidth) and a 256 point DFT is computed once every 5 msecs to obtain a wide-band spectrum. From the wide-band spectrum, a magnitude squared spectrum is derived. Energy is computed on the magnitude squared spectrum in a tapered frequency range of 300 to 4000 Hz, covering the sonorant region of the spectrum. The fundamental frequency is computed every 5 msec using the method developed by Seneff [22].

2.1.1 Duration

The durations of the vowel regions are measured from the boundaries established in the phonetic transcription. Sixty-five percent of the vowels perceived as stressed in the listening test were maximal in duration across the word. Measuring the duration of the sonorant region of each syllable within the word yields comparable

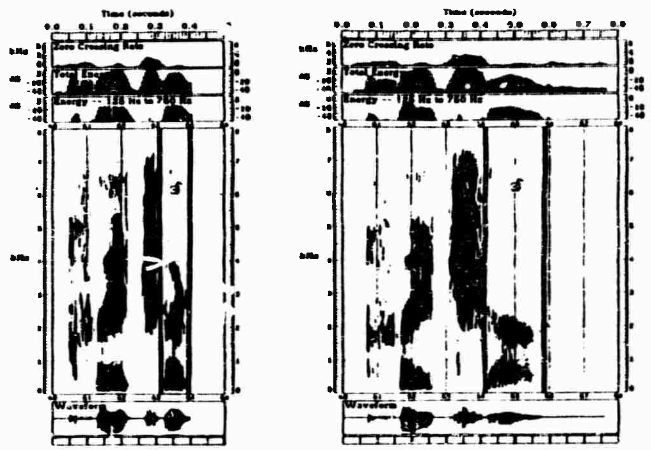
[1] Features in this thesis are defined to be acoustic attributes as opposed to distinctive features [3]

results. In this case, 63% of the stressed syllables have the maximum duration value within the word.

As discussed by Klatt [11], prepausal lengthening increases the duration of the word final vowel by as much as 30% in phrase final position. Since the words in our database are spoken in isolation, Klatt's results may present a lower bound of the effect of prepausal lengthening. In the example shown in Figure 2-4, a wide-band spectrogram of the word *creature* is shown first spoken in a carrier phrase and then spoken in isolation. The word final vowel duration for the word in a carrier phrase is only 50% of that of the vowel in the isolated case. For the 35% of the words where the stressed syllable does not have a maximum duration, close to 75% of those cases were due to a maximum duration value on the final unstressed vowel. In order to compensate for this effect, the duration of the vowel or sonorant region in syllable final position is reduced to 60% of its original duration. In this case, 90% of the stressed vowels are maximal in duration across the word. From these results, duration appears to be a robust acoustic cue for stress in isolated words after the effect of prepausal lengthening has been compensated.

Our duration results are comparable to those of Fry's experiment which claimed that the perception of stress is influenced more by duration than intensity. In his experiment, as the duration ratio of the two vowels within two syllable words is increased from 2 to 5, the percent of words perceived to have stress on the longer vowel increases from 40% to 90%. Similarly, his manipulation of intensity ratios results in stress perception increasing from 50% to only 70% with increasing intensity ratios. In Lieberman's acoustic measurements, he found that 66% of the stressed syllables in his database had a longer duration than the other syllables within the words (without compensation for prepausal lengthening).

In summary, duration seems to be a valuable feature to incorporate into a stress recognition system as it is relatively easy to measure and is an acoustic correlate of



"CREATURE"

extracted from a carrier phrase

"The CREATURE destroys people."

"CREATURE"

spoken in isolation

Figure 2-4: As an example of the effect of prepausal lengthening on the word final sonorant, the word *creature* is spoken (a) in a carrier phrase and (b) in isolation.

lexical stress. There are two additional points of interest in this duration study. First, one can account for the effect of prepausal lengthening. Second, the measurement of duration in the sonorant region as opposed to the vowel region is adequate as the results do not substantially deteriorate.

2.1.2 Energy

Previous work provides evidence that stressed vowels display an increase in energy in the lower half of the spectrum. To quantify this, the energy in the frequency range from 300 to 4000 Hz is computed for each utterance in the database. The average energy from samples i to j, $\langle Energy \rangle_{ij}$ is computed as follows.

$$\langle ENERGY \rangle_{ij} = \log \sum_{n=i}^{j} E[nT] \Big/ (j-i)$$

where $E[nT]$ = energy value at time nT

and $(j-i)$ = the length of time the energy is being averaged

T = 5 msec

An average energy is computed for each vowel or sonorant region for each word in the database. The maximum of the average energies of the vowel regions across the word corresponds to the perceived stress 84% of the time. The maximum average energy of the sonorant regions produces comparable results as 79% of the stressed syllables have an energy maximum.

Similar to duration, energy appears to be a strong correlate of stress whether measured in the vowel region or in the sonorant region of the syllable. Its average value over the sonorant region illuminates stress information in over three-fourths of the observed data.

2.1.3 Fundamental Frequency

The extraction of fundamental frequency (F0) is based on an enhanced waveform which is the sum of the rectified and amplitude-compressed outputs of a filter bank covering the frequency range from 200 to 2500 Hz. With this method, developed by Seneff [22], the waveform displays strong periodicity at the fundamental allowing an Average Magnitude Difference Function of the waveform and a voicing decision to produce an accurate F0 contour. Similar to the energy measurements, the average F0 is computed in each vowel and sonorant region of the word. The maximum value of the F0 average in each syllable corresponds to the stressed syllable in 61% and 55% of the cases, respectively. If the F0 values are first smoothed and then the maximum value extracted from each region without averaging, the previous results improve by 15%. This shows that in over 70% of the words stress produces peaks in F0 rather than raises the overall F0 average.

The relatively inconsistent results of F0 on a stressed syllable as compared to duration and energy can be partially attributed to the nature of words said in isolation. Without a carrier phrase, words are often-pronounced with an increase in F0 in the final syllable independent of the position of primary stress. In Lieberman's work, he associated maximum values in F0 with stress and found that 90% of the stressed syllables exhibit such a F0 maximum. There may be several reasons for the discrepancy between his results and ours. Besides different means of pitch tracking and smoothing, Lieberman extracted his words for analysis from continuous speech. This helps prevent a *sing song* effect that distorts the F0 values. Words read in isolation from a list are more susceptible to such rhythmic effects.

2.2 Lexical Constraints Provided by Stress

The acoustic study establishes the general feasibility of extracting correlates for stress from any isolated word. Specific recognition systems have been designed that would benefit from prosodic information. Such a system proposed by Huttenlocher [10] is a speaker independent, isolated word system with a vocabulary of 20000

words. In his system, the speech signal is crudely segmented into broad classes. Word candidates are proposed by matching the derived phonetic classification of the signal against those stored in the 20000 word lexicon. As discussed in Chapter I, prosodic information in addition to segmental provides additional lexical constraint. In particular, the segmental information around stressed syllables is more constraining than that around unstressed. Thus, in addition to illuminating regions of acoustic reliability, stress may provide strong lexical constraints for the specific task of large vocabulary isolated word recognition by lexical access.

In this section, the lexical constraints provided by stress alone are investigated for a 20000 word lexicon, the Merriam-Webster Pocket Dictionary (MPD). Each word in the lexicon is associated with one pronunciation baseform. If the number of syllables in each word is determined by the number of vowels or syllabic consonants in the pronunciation, then the distribution of lexicon by number of syllables is shown in Figure 2-5a. The distribution reveals that 75% of the words consist of two, three, and four syllables. This is assuming that all the words are equally weighted or equally used in the English language. A more realistic approach would be to weight the words by their frequency of usage[2]. This captures the realistic distribution of the 20000 words were someone to choose words at random to input to a recognition system. Monosyllabics can be discarded mainly due to the fact that they do not provide an interesting case for stress as they only have one level of stress. Without monosyllabics, there are over 16000 polysyllabic words remaining. Incorporating frequency of usage of the words in the English language as specified by the Brown corpus redistributes the polysyllabic sublexicon such that 97% of the words fall into the two, three, and four syllable category. (See Figure 2-5b.) These results indicate that the focus of this study should be on the two, three, and four syllable words.

[2] Words from the lexicon that do not appear in the Brown Corpus are arbitrarily given a frequency count of one

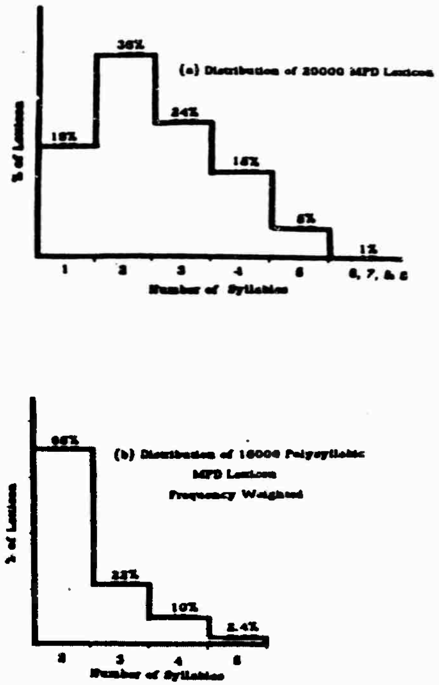


Figure 2-5: Distribution of (a) MPD 20000 Lexicon with Equal weighting and (b) of MPD 16200 Polysyllabic Words with Frequency weighting

A subset of the polysyllabic words, namely the 15000 two, three, and four syllable words, in MPD are mapped into stress pattern classes, where there are three levels of stress. To label each word with its appropriate stress pattern, the following guidelines are used. The number of syllables equals the number of vowels and syllabic consonants in the baseform. Primary stress, as marked in the baseforms, establishes the stressed syllable. Schwa's in the pronunciation are considered as

reduced. The remaining syllables are marked as unstressed. For example, [SUR] is the class of three syllable words, such as *calendar*, with stress on the first syllable and a reduced vowel in the last syllable.

The first experiment assumes that the stress level for every syllable in the word is known. The word *calendar* would be associated with the stress pattern class [SUR] as stated above. This represents an ideal case in that the stress quality of each segment is known. When the entire lexicon is mapped into these stress pattern classes, the expected value of the class size is computed as discussed in Chapter 1. To render the classification most realistic, frequency of usage of the words is incorporated (as previously discussed). Thus, the probabilities for each class are adjusted to reflect the frequency weighting. Two additional smaller lexicons are also classified and evaluated to ensure that the classification results are general and not specific to a particular lexicon. The results, shown in Table 2-1, reveal an expected class size that is 19% of the 15000 polysyllabic words with a maximum class size of 22%. These results are encouraging since, if one were given the stress pattern for a particular word, one could expect to reduce the number of possible word candidates immediately from 15000 to less than 3000 with no segmental information. Frequency weighting of the words decreases the expected class size by 8% mainly because the largest stress pattern classes also have the greatest frequency of usage.

These results are an indication of the constraining power of the complete stress pattern information in combination with a knowledge of the number of syllables. However, a classification such as this may not be robust as it demands precise knowledge of all reduced and unstressed segments, as well as primary stress. These are stringent requirements and would surely be prone to error in a practical implementation. Therefore, further experiments are performed where less knowledge is assumed. The original experiment forms a basis against which to compare the constraining power of other classification criteria.

Lexicon Size	15C00	2000	1100
Expected Class Size	2916 19%	303 15%	148 13%
Maximum Class Size	3360 22%	801 31%	297 26%
Frequency Weighted Expected Class Size	1688 11%	335 17%	175 15%

Table 2-1: Results of Stress Pattern Classification of Lexicons Given All Stress Information

The next experiment investigates the constraints provided by stress pattern classes that rely solely on the number of syllables and the location of primary stress. All other syllable positions carry no stress level information. The word *calendar* would be a member of the class [S••] where the unstressed and reduced syllables are merely placeholders. This relaxed criterion presents a more achievable means of classification assuming an ultimate dependence on the acoustic signal. The results in Table 2-2 illustrate an expected class size of 21% of the large lexicon with a maximum class size of 37%. These are promising results as the difference between the expected class size given all stress information, 19%, and the expected class size given only the position of primary stress, 21%, is small. In other words, if all the stress level information cannot be extracted, it is still desirable to locate the position of primary stress as it provides the majority of lexical constraint. These results are consistent across all lexicon sizes as well. Frequency weighting of the words increases the relative difference in the expected class sizes more significantly in the large lexicon. These results suggest that the location of primary stress is a practical goal that may provide adequate lexical constraint.

Lexicon Size	15000	2000	1100
Expected Class Size	3180 21%	347 17%	293 26%
Maximum Class Size	5547 37%	938 47%	466 41%
Frequency Weighted Expected Class Size	3684 24%	630 31%	332 29%

Table 2-2: Results of Stress Pattern Classification of Lexicons Given Only Primary Stres

The next experiment considers the constraints imposed if the classification consists of the primary stress position and the position of one reduced vowel with all other syllables wildcarded. In this case, we are assuming that there may be enough information to classify additional syllables but perhaps not the entire word as was done in the first experiment. The results in Table 2-3 show that the additional reduced information constrains the expected class size to 8% of the large lexicon with a maximum class size of only 15%. The lower values of these expected class sizes can be misleading when compared to the results of the original experiment, shown in Table 2-1, since not all words contain reduced vowels in their lexical representations even though they may be reduced in natural speech. Secondly, the classes are not mutually exclusive as a word may contain more than one reduced segment unlike primary stress. Yet, information concerning the location of primary stress and some reduced segment does provide some additional constraint.

Lexicon Size	15000	2000	1100
Expected Class Size	1160 8%	129 6%	73 7%
Maxim. m Class Size	2187 15%	337 17%	171 16%
Frequency Weighted Expected Class Size	1295 9%	127 6%	72 7%

Table 2-3: Results of Stress Pattern Classification of Lexicons Given Primary and a Reduced Segment

Finally, we consider the case where no stress information is available. Instead, only the number of syllabies is known. This is important as it differentiates the constraints provided by stress and those provided by the number of syllables. Again, the three syllable word *calendar* would be classified as [• • •]. As shown in Table 2-4, the large lexicon has an expected class size corresponding to 37% and 41% of the non-weighted and weighted lexicon, respectively. Knowing the syllable content strongly constrains the lexicon but the additional information of only the stress syllable can reduce the expected class size by almost half. Thus, finding the number of syllables and the position of stress may be a sufficient means of lexical constraint as well as a practical goal for implementation.

In summary, stress provides enough lexical constraint to benefit the task of large vocabulary recognition. It is adequate, as well as more practical, to identify only the regions of primary stress and the syllables as the majority of constraint is provided by knowledge of the number of syllabies and the location of the stressed syllables.

Lexicon Size	15000	2000	1100
Expected Class Size	5550 37%	1029 51%	568 52%
Maximum Class Size	7180 48%	1328 66%	737 67%
Frequency Weighted Expected Class Size	6192 41%	1073 54%	613 56%

Table 2-4: Results of Classifications of Lexicons Given Only the Number of Syllables

2.3 Chapter Summary

The major points of this chapter are:

- There is evidence that lexical stress exhibits physical correlates that can be extracted from the speech signal. Duration, energy, and fundamental frequency are potential features for stress determination;

- The identification of acoustic features for stress does not rely on locating exact vowel boundaries since locating the sonorant regions of the syllables does not deteriorate the stress information;

- It may be the proper combination of these features and not their isolated performance that provides an effective stress detection algorithm;

- The task of large vocabulary speech recognition can be reduced by the addition of this prosodic information;

- Stress information provides a great deal of lexical constraint for this specific recognition task;

- Knowing the number of syllables and the position of the primary stress, as opposed to the entire stress pattern, provides adequate constraints as well as a practical goal.

Chapter Three

Stress Recognition Implementation

3.1 Introduction

Stress information is valuable to speech recognition systems due to its ability to provide pointers to regions of the speech signal that are acoustically robust. In addition, the lexical studies reported in this thesis demonstrate the constraining power of stress information for the particular task of large vocabulary isolated word recognition. Focusing on lexical stress reduces the task to isolated words instead of continuous speech, thus eliminating the complex effects of sentential stress.

The initial acoustic studies on phonetically transcribed data are evidence for the potential to extract acoustic features that illuminate stress information. Algorithms that provide a detailed phonetic segmentation from the speech signal are often prone to error due to the inherent inter- and intra-variability of speech. Therefore, in a practical implementation, it should not be assumed that accurate detailed phonetic information is available for stress determination if the information is to be derived solely from the speech signal.

More reliable segmentation schemes are based on classifying the speech signal into broad phonetic classes [10] [17]. The advantage of broad classes, as discussed previously, is that the classes are acoustically robust and relatively invariant with respect to other classes and speakers. The accuracy in labeling segments of the speech signal as sonorant is much greater than individually labeling vowels, liquids, glides, and nasals. Once the initial segmentation labels regions of the speech signal that are sonorant, it may be necessary to examine these areas in greater detail for additional syllable boundaries. This stage is more difficult as it is harder to distinguish sounds within the same class than it is to distinguish sounds across classes, such as differentiating a sonorant from a fricative. As discovered in the acoustic study of Chapter 2, deriving stress features from the sonorant regions of the syllables as opposed to only the vowel regions does not appreciably distort the features' manifestation of stress. Thus, a more practical and accurate means of syllable boundary detection is afforded as it is extremely difficult to extract a vowel from its surrounding semivowels or nasals.

Once the sonorant regions for each syllable are identified, features for stress can be extracted for each syllable. A comparison of the features derived for each syllable could establish the appropriate stressed syllable and subsequently illuminate regions of acoustic reliability. In order to demonstrate the lexical constraints of stress information in a specific recognition scheme, the stress pattern can be mapped into a large lexicon similar to the previous lexical studies. In the remainder of this chapter, the actual stress recognition implementation is discussed in detail.

3.2 System Outline

The system designed in this thesis is outlined in Figure 3-1. The input is an isolated word of American English. The first stage is the front-end processor which consists of two parts. First, an initial broad segmentation [17] provides pointers to sonorant regions and their boundaries. The second part examines these regions in greater detail for additional syllable boundaries. Possible syllable boundaries encountered are intervocalic semivowels, as in the word *yellow*, or a sequence of two vowels, as in the word *create*. Thus, more difficult further segmentation of the sonorant regions is performed in this part of the system. The front-end processor eliminates the need for any transcription to be provided as all the phonetic information is derived directly from the acoustic signal.

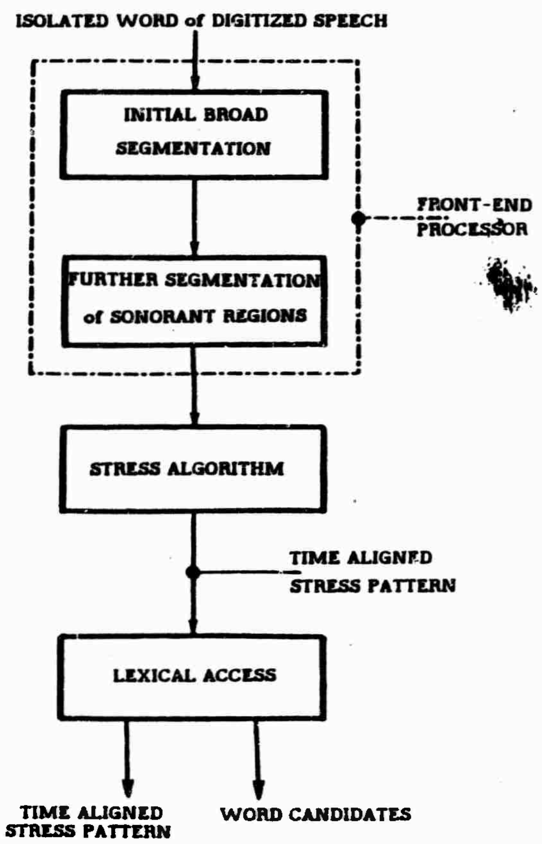

ISOLATED WORD of DIGITIZED SPEECH
INITIAL BROAD SEGMENTATION
FRONT-END PROCESSOR
FURTHER SEGMENTATION of SONORANT REGIONS
STRESS ALGORITHM
TIME ALIGNED STRESS PATTERN
LEXICAL ACCESS
TIME ALIGNED STRESS PATTERN
WORD CANDIDATES

Figure 3-1: Stress Recognition System Outline

Once syllable regions are located, features for stress can be measured in each region. The features extracted consist of duration, energy, fundamental frequency, and spectral change. Each syllable is then associated with a feature vector. Judgment of stress is made by a comparison of feature vectors within a word. Thereby, the judgment remains a relative decision within the given word as opposed to a global one across all data. The stress algorithm assigns each sonorant region to be stressed, unstressed, or reduced (schwa), with only one syllable receiving a stressed label corresponding to the primary stress. For cases in which the acoustic cues are ambiguous, or two syllables appear to have proximate feature vectors, a second choice for a stressed label is assigned. These cases may correspond to a secondary stress such as the first syllable in the word *Massachusetts*. The output of the stress detection is a time-aligned stress pattern.

To complete the system, lexical access is then performed based on the polysyllabic words of MPD, where each word is associated with its baseform pronunciation. However, in order to account for syllable reduction, transformation, or deletion, the words in the lexicon are expanded by a set of phonological rules. For example, even though the word *desire* has a two syllable representation in the original dictionary, /dəzɑ'r/, one pronunciation of it may be three syllables, /dəzɑ'ɝ/. When the polysyllabic words in MPD are expanded by such phonological rules, the lexicon is expanded by 75%. Thus, the final output of the stress recognition system is a time-aligned stress pattern and an equivalence class of word candidates that could be associated with the derived stress pattern. In the remainder of this chapter, the individual components of the system will be discussed in greater detail.

3.3 Front-end Processor

The front-end processor consists of two stages: an initial broad segmentation and a more detailed examination of the sonorant regions. The first stage by Leung [17] segments the speech signal into broad classes using a non-parametric pattern classifier. The structure is that of a series of binary classifiers arranged in a decision tree. Features for each classifier are selected based on the decision being made. The decisions at each node are made on the feature vectors by a K-Means clustering algorithm, using a Euclidean distance metric. Thus, on a frame by frame basis (every 5 msec) the speech signal is segmented into broad classes: sonorant, obstruent, voiced obstruent, silence, nasals, and voice bars. Regions that cannot be classified remain unlabelled.

The sonorant regions of the word are extracted from the initial broad segmentation. These regions are then further examined for additional syllable boundaries. In the broad segmentation, intervocalic nasals may not always be located. Also, the segmentation does not attempt to identify semivowels. In a word such as *Massachusetts*, the syllables are separated by obstruents and are readily segmented as shown by the wide-band spectrogram in Figure 3-2. The lines drawn correspond to the sonorant regions which, in this case, also correspond to each of the syllables. Yet for the word *yellow*, shown in the same figure, the entire word consists of sonorants. The semivowel, /l/, makes the two syllables harder to separate. In some cases, there may be no intervocalic consonants but only a sequence of vowels, such as in the word *anxiety*. The task of separating the sonorant portions of the syllables can be broken down into subproblems of increasing complexity.

The easiest cases in which to separate the syllables are words with obstruents between syllables, such as *Massachusetts*. The words require no further segmentation after the initial broad segmentation as each sonorant region corresponds to a syllable. The next less difficult cases are intervocalic nasals,

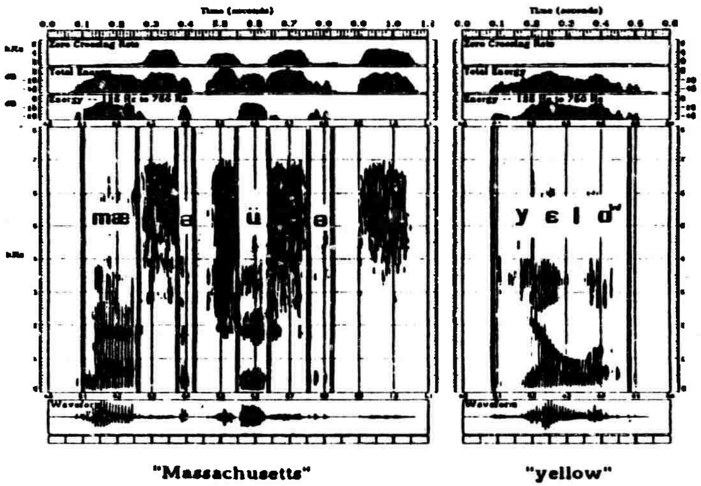


Figure 3-2: Wide-Band Spectrograms of the words *Massachusetts* and *yellow* showing the increasing difficulty locating syllable units

/mnG/. Nasals in this context usually exhibit a significant decrease in low frequency energy as well as an appreciable amount of spectral change, i.e., the amount of change in the spectral shape with time. As seen in Figure 3-3a, the

energy from 1000 Hz to 3000 Hz shows a substantial decrease during the nasal portion of the word *flannel*. The energy values corresponeding to the sonorant portion of the word are darkened for emphasis in Figure 3-3. Robust dips in the energies in the regions of the first formant (F1), the second formant (F2), and the third formant (F3) are used to locate syllable boundaries.

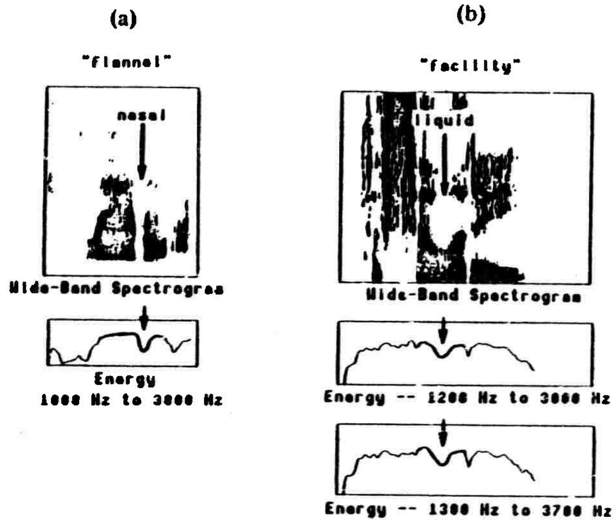


Figure 3-3: Examples of energy contours that capture intervocalic
(a) nasals as in the word *flannel* and
(b) liquids as in the word *facility*

The intervocalic semivowels, /l, w/, are more difficult to locate than intervocalic nasals. Typically, there is a decrease in energy in the F2 and F3 regions. However, the spectral change is fairly smooth and gradual. Thus, semivowels have fewer features that distinguish them from vowels. As shown in Figure 3-3b of the word *facility*, gradual dips in the energies from 1200 Hz to 3000 Hz or from 1300 Hz to 2700 Hz may be a cue for intervocalic semivowels. For both the nasals and semivowels, the normalized energy waveforms are characterized such that robust transitions are captured. Setting stringent thresholds for what should be considered as *high* or *low* energy levels allows these intervocalic consonants to be characterized as a *high-low-high* transition.

A more difficult task is that of finding a syllable boundary between adjacent vowels. The difficulty of this is compounded by the fact that determining the boundary between two vowels is not entirely obvious even from visual examination of a spectrogram. However, the majority of *vowel-vowel* sequences in the MPD lexicon constitute a vowel quality transition such as high to low or front to back.

In summary, the initial broad segmentation delineates sonorant regions that may contain *vowel-sonorant* sequences. In order to find the syllable regions, further processing that focuses on contextual information and the temporal variation of the signal should be employed.

Theoretically, if the formant values were known, then a decision could be made as to whether there was a suitable change in formant position with time. In some cases, the formants are well separated and stable with time. One could successfully track their movement with such measures as center of gravity of a spectral slice or spectral peak picking. However, accurate formant tracking remains elusive as the formant movement is highly context dependent. Some of the difficulties encountered include nasal formants, merging formants, as well as weakened formants [14]. In some cases, tracking the position of the formants is even difficult by eye. In addition, most effective formant tracking algorithms are computationally expensive. The information sought by the detailed segmentation process is a relative change in vowel quality within a sonorant region as opposed to vowel identification. Formant tracking seems to be overkill for such a task. An appropriate mechanism should be simple yet able to capture the change in energy distribution corresponding to a change in vowel quality.

One attempt to illuminate the change in energy concentration of the formants is a center of gravity measure, or the first moment of each spectral slice, which displays the amount and position of spectral energy with time for a specified frequency range. The center of gravity, COG, essentially weights the spectral slice by a linear window and sums over the weighted spectral values as shown below.

$$\mathrm{COG} = \sum_{f=F_1}^{F_2} W(f)\, S(f)$$

where $W(f)$ = linear weighting window

$S(f)$ = spectrum value at frequency f

F_1, F_2 = frequency range

There are two disadvantages to a linear spectral weighting window. First, the sensitivity of the weight can be examined by computing its first derivative shown in Figure 3-4. The derivative is constant except at the edges where the discontinuity of the edges makes the weighting window very sensitive at the endpoints of the frequency range. In other words, changes in the spectrum near the endpoints will be inflated over those in the stable part of the window. This may not reflect the relevant changes in the spectrum but emphasize the irrelevant ones.

A second disadvantage of the linear weighting window is that it emphasizes high frequency information over low frequency which may not always be appropriate for speech. For example, a center of gravity from 0 to 1000 Hz may attempt to capture

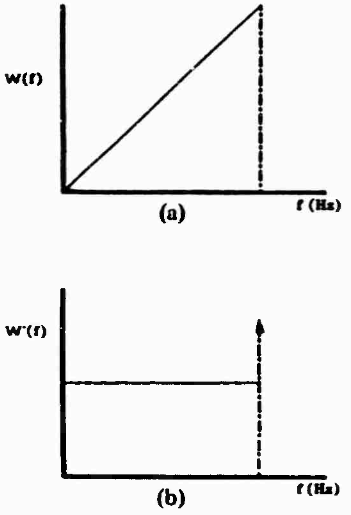


Figure 3-4: Center of gravity measures involve a (a) linear spectral weighting window with a (b) sensitivity function manifested by its derivative

the position of F1. A high vowel to low vowel transition could also be captured in this way. However, the linear window would weight the frequency range above 500 Hz (corresponding to a low vowel) greater than the range from 0 to 500 Hz (corresponding to a high vowel). The center of gravity information, then, may not be appropriate due to the nature of the linear weighting window.

An alternative to the above center of gravity measure is to choose a weight that is sensitive at crucial points and, more importantly, incorporates better speech specific knowledge to weight the spectrum. For example, in order to make a front-back

vowel distinction, a specific weighting window could be designed for that task. In the case of a back vowel, F1 and F2 are close together as compared to F2 and F3. However, for a front vowel F2 and F3 are relatively close. To capture the energy concentration as it moves from a front to back vowel, the sinusoidal weighting window in Figure 3-5a replaces the linear one. The energy in the spectrum below 1500 Hz receives a positive weight while the energy above 1500 Hz receives a negative weighting. If one examines the sensitivity of the sinusoid window by the first derivative as shown in Figure 3-5b, it is very sensitive at the crossover point, 1500 Hz. As a neutral vowel is considered to have formants around 500, 1500, and 2500 Hz, this window then reflects changes of the vowel from neutral position to front or back positions. When a vowel changes from back to front (or vice versa), the second formant may cross through the 1500 Hz mark where appropriately the window is sensitive. Thus, the sinusoidal weighting emphasizes relevant transitions. This pseudo center of gravity measure with the specific window would be more negative for front vowels and more positive for back vowels. A vowel quality transition could be realized as a change in the pseudc center of gravity from positive to negative (or vice versa).

A similar window, is used to make a high-low distinction. A high vowel has a low F1 (below 500 Hz) while a low vowel has a relatively high F1 (above 500 Hz). The window is very sensitive at the crossover point, around 500 Hz, through which F1 moves in a high to low or low to high transition. Energy concentrations below 500 Hz are given a positive weight while those above 500 Hz receive a negative weight. If the pseudo center of gravity using this weighting window changes from positive to negative, the vowel quality has probably changed from high to low.

The sum of the weighted spectral components at each point in time yields a time domain waveform that can be readily characterized by its transitions from positive to negative or vice versa. In the more detailed segmentation, this is the means of

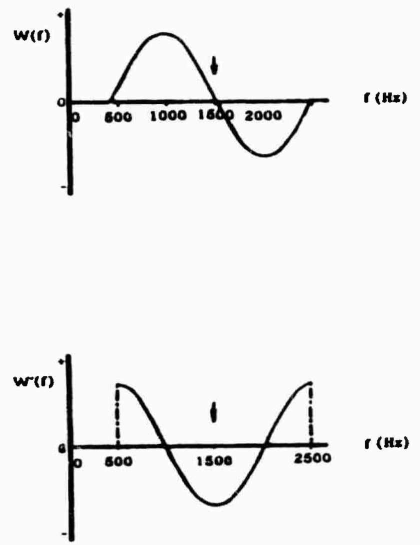


Figure 3-5: Spectral weighting window that emphasizes changes in vowel quality from a front vowel to a back vowel

locating a change in vowel quality that may be associated with adjacent vowels. Thus, a syllable boundary can be placed where the output of pseudo center of gravity makes a significant change in sign.

Thresholds enable the waveform to be characterized as *high* or *low*. A vowel-vowel sequence is then recognized as a *high-low* or *low-high* transition as shown for the word *society* in Figure 3-6. In this case, two spectral weights of a period of a cosine wave in the frequency ranges 900 to 3100 Hz and 500 to 2500 Hz, both show a robust transition as the vowel region changes in quality.

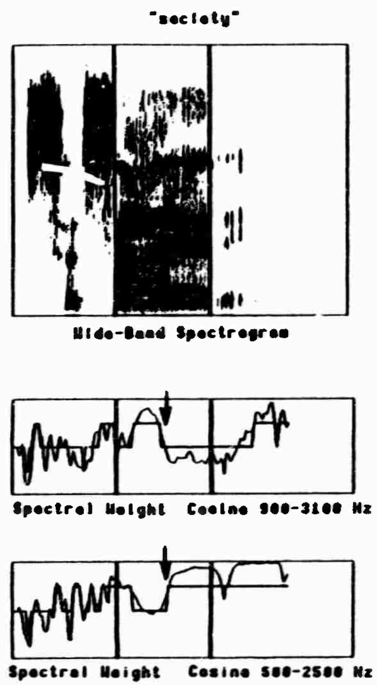
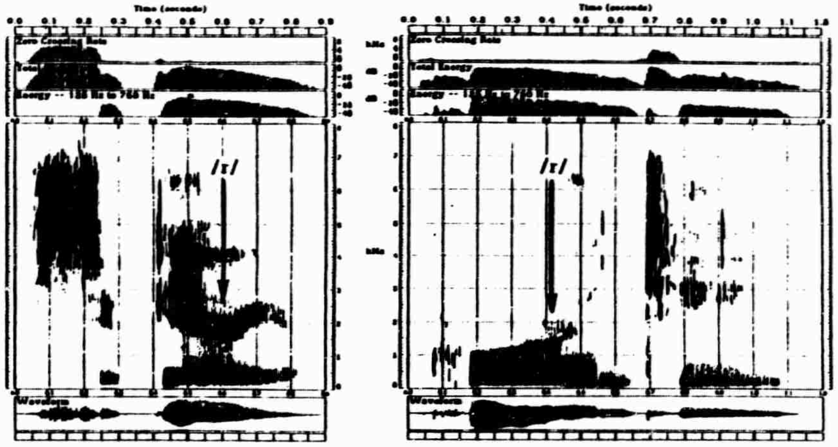


Figure 3-6: Spectral weighting window that emphasizes changes in vowel quality for the word *society*

To avoid placing a syllable boundary in the middle of a diphthong, a durational constraint is used. This states that if the first segment is over 30% longer than the second segment, and the second segment is not reduced (i.e., less than 60 msec), then the region may be a diphthong and the syllable boundary should be discarded. This constraint is necessary because a problem encountered with the spectral weighting is that it captures the formant movement in diphthongs as well as adjacent vowels. For example, there is pronounced formant movement in the diphthong /Y/. A sequence of two vowels and a diphthong may both have robust changes in energy distributions that can be readily captured by the spectral weighting.

The final difficult class of intervocalic sonorants to locate is the semivowels /r/ and /y/. The semivowel /y/ does not present a problem as it is usually considered to be an off glide of the vowel. It occurs intervocalically in less than .1% of the words in the MPD. However, intervocalic /r/'s occur quite often and should be incorporated in the further more detailed segmentation process. The striking acoustic characteristics of an /r/ are usually a third formant just below 2 kHz and a low F1. There is a prominent dip in F3 when the /r/ is surrounded by front vowels such as in the word *superior*. (See Figure 3-7.) Yet, the presence of an /r/ can be quite disguised in some environments, such as the word *quarantine* shown in the same figure. The difficulties involved in locating intervocalic /r/'s are similar to the problems associated with a vowel-vowel sequence. The /r/ sound is the most *vowel-like* semivowel as there may be little or no movement or weakening of the formants. Looking for dips or extreme changes in distribution of energies may not illuminate information regarding the presence of an /r/. However, the typical concentration of energy around 2 kHz and below the F1 region indicates that spectral weighting may again be applicable in labeling segments as roughly *r-like* or *not r-like*.

A specific window to weight the spectrum for an /r/, shown in Figure 3-8, reinforces sounds that have energy in the low frequency and 2 kHz regions. Other regions of the spectrum are given a negative weighting to avoid confusion of /r/'s with vowels such as /i/ or /æ/. The output of the spectral weighting is characterized as being *high* or *low*, i.e. *r-like* or not. A syllable boundary corresponding to an intervocalic /r/ may be placed where there is a transition of low to high to low. Figure 3-9 shows an example of the spectral weighting (or the pseudo center of gravity) for the word *experience*. In this example, the /r/ is



"superior" "quarantine"

Figure 3-7: Wide-band spectrograms of the words (a) *superior* and (b) *quarantine* showing difference in difficulty in locating intervocalic /r/'s

captured in its intervocalic context as a transition in the parameter from low to high to low. A syllable boundary is placed in the center of the high segment as shown by the arrow in the figure.

In the further segmentation, there are several spectral weights used to capture vowel quality changes and intervocalic /r/'s. Using several weights for the vowel transitions may alleviate interspeaker variability. Cosine weights from 200 to 2800 Hz and from 500 to 2500 Hz distinguish front vowels from back vowels. A half cosine from 50 to 1050 Hz separates the F1 regions so that high and low vowel quality transitions can be captured. Finally, the /r/ weighted window emphasizes the spectrum around 2000 Hz and below 500 Hz.

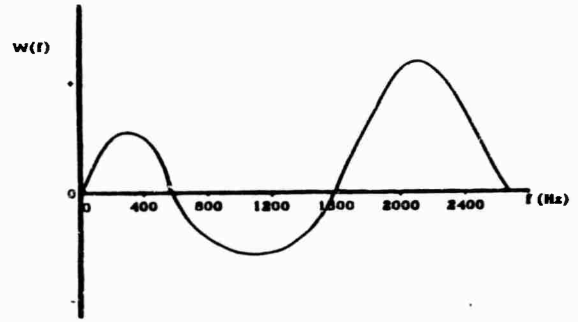


Figure 3-8: Spectral weighting window designed to emphasize *r-like* spectral shapes

The following gives an example of the work done by the front-end processor. Figure 3-10a shows a spectrogram for the word *anxiety*. Directly below the spectrogram is the output of the initial broad segmentation where the label "S" represents the sonorant regions. Figure 3-10c and 3-10d are energy waveforms in the frequency ranges 1000 Hz to 3000 Hz and 1300 Hz to 3700 Hz, respectively. The characterization of these waveforms shows no evidence for a syllable boundary. In Figure 3-10e, a spectral weighting by a cosine from 500 to 2500 Hz shows a transition that is identified by the front-end as a syllable boundary. Both segments on either side of the new boundary have enough energy in the range 750 Hz to 4000

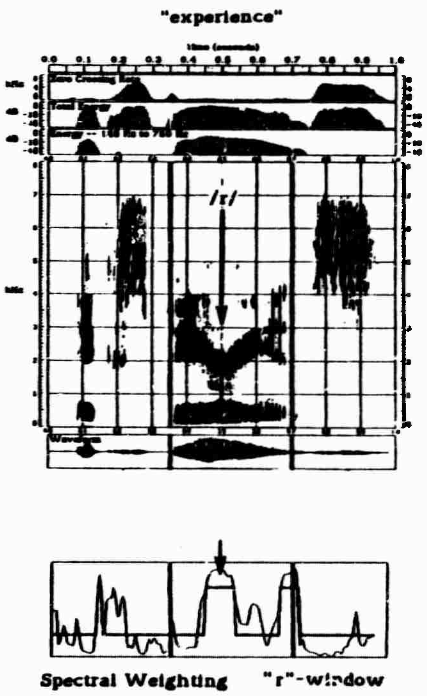


Figure 3-9: Example of spectral weighting for an intervocalic /r/ in the word *experience*

Hz to qualify as syllables. Such an energy check ensures that pre- or post-vocalic consonants are not labeled as a syllable unit. The final output of the front-end processor is transcription that identifies the sonorant portions of the syllables and their time boundaries. The bottom of Figure 3-10 displays this information.

In summary, the front-end processor consists of two major portions. The initial crude segmentation identifies sonorant regions of the speech signal but does not

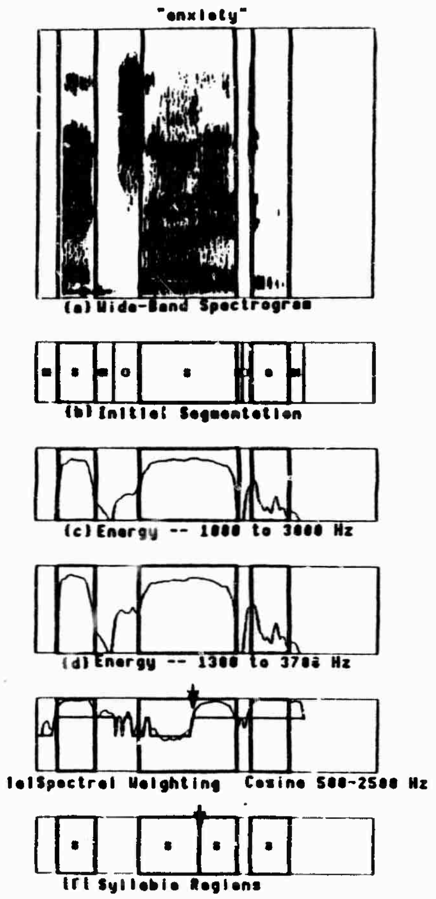


Figure 3-10: Example of the work done by the front-end processor for the word *anxiety*

necessarily uniquely identify the syllables. The further segmentation attempts to locate several classes of sounds, usually in an intervocalic context, to establish syllable boundaries when consonants can not do the separating. These classes in increasing order of difficulty are nasals, /l/ and /w/, vowel-vowel sequences, and /ɾ/. Charecterizing energy and spectrally weighted contours allows transitions for these sounds to be located as possible syllable boundaries. When a boundary is located, the amount of energy in the lower half of the spectrum is checked to ensure that the segment should be considered a vowel.

3.4 Stress Algorithm

3.4.1 Parameters

Once sonorant syllable regions are established, features that illuminate stress information are obtained for each syllable. From the acoustic study of Chapter 2, duration, energy, and fundamental frequency are potential acoustic correlates of stress. However, individually, they achieved only an 87% performance rate in the best case. This suggests that a combination of the parameters may attain a higher performance. In addition, stressed syllables, with the exception of diphthongs, seem to be more steady state or spectrally stable with time. For this reason, a measure of spectral change is included as a parameter for stress. Thus, the following parameters are extracted from the original waveform: duration energy, fundamental frequency, and spectral change. Duration is obtained from the boundaries established by the front-end processor. Energy is computed in the frequency ranges 400 Hz to 5000 Hz and 1200 Hz to 3300 Hz. Fundamental frequency is computed as discussed in Chapter 2 and then passed through a smoothing mechanism. Spectral change is based on the difference in energy values over several adjacent frames of data and will be discussed in greater detail below.

After the parameters are extracted, they are normalized. Each parameter has a predetermined range which is linearly mapped into the normal range as shown in Figure 3-11. The range for each parameter is chosen to cover all reasonable values for that parameter. In other words, from the observation of over 1000 utterances with the above parameters computed, there are no parameter values that fall outside their range. If a parameter value did not fall into its predetermined range, it would probably be due to a front-end processing error. These are very conservative range values but are tuned to the specific segmentation and parameter computations. Each syllable is then associated with five feature values, each ranging from 0 to an arbitrary number x. The parameter extraction is discussed in detail below.

3.4.1.1 Duration

The duration is based on the sonorant region boundaries obtained from the front-end processor. Since the segmentation is based on parameters computed once every 5 msec, each boundary established could be off by as much as ±5 msec. As demonstrated in the initial acoustic study of this thesis, the final syllable duration must be compensated for prepausal lengthening. As discussed by Klatt [11], the amount of word final vowel lengthening depends on its following context. Since the segmentation provides some broad contextual information, the amount of compensation can be varied according to the context following the final sonorant region. Specifically, the conditions are as follows.

Context following word final sonorant	Percent Duration Diminished
Obstruent	15%
Voiced Obstruent	20%
Nasal or unlabeled	30%
silence	50%

If the word final sonorant is followed by an unvoiced obstruent, there is less lengthening than if it were followed by a voiced obstruent. If the final sonorant is

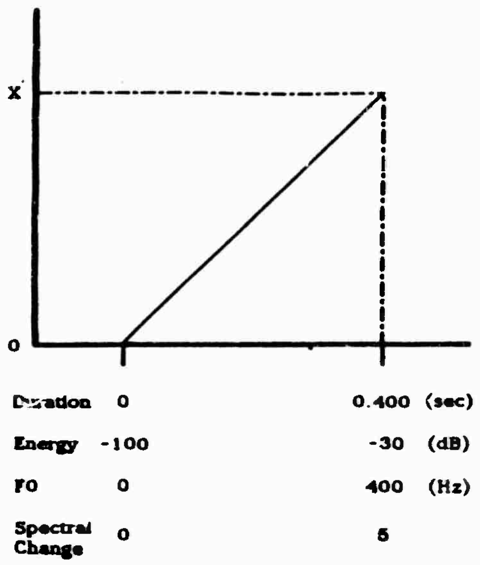


Figure 3-11: Linear mapping of stress parameters into arbitrary range for normalization

also the last segment of the word, it will undergo a maximum amount of lengthening. Thereby, the percentage of compensation is adjusted according to the final context derived from the initial segmentation, ranging from a 10 to 40% decrease in duration.

3.4.1.2 Energy

The energy is computed in two frequency ranges of the spectrum: 400 Hz to 5000Hz; and 1200 Hz to 3300 Hz. The former range covers the entire sonorant region. The latter is essentially the F2-F3 region and is intended to deemphasize syllabic consonants and voice bars as they are not usually recipients of primary stress. In each syllable region, the average of both energies is calculated as shown in Chapter 2. The logarithm of the averages is then computed. Thus, each syllable is associated with two log energy values.

3.4.1.3 Fundamental Frequency

In order to locate valid maxima, the F0 waveform needs to be smoothed. A global smoothing mechanism finds the average and standard deviation of all nonzero values. Each point along the waveform is then examined for continuity and distance from the average. Stray points are reset to zero. This global smoothing remedies pitch doubling as well as voicing errors so that a meaningful maximum F0 value can be associated with each syllable region.

3.4.1.4 Spectral Change

Stressed segments usually exhibit more spectral stability than unstressed. Therefore, spectral change [17] is measured as a difference in sonorant energy values over time. The sonorant energies consist of a bank of energies, namely 0 to 300 Hz, 300 to 600 Hz, 600 to 900 Hz, 900 to 1200 Hz, 1200 to 1500 Hz, 1500 to 1800 Hz, and 1800 to 2100 Hz, all normalized by the total energy in the spectrum. The spectral change measure, S[n], is shown below.

$$S[nT] = \max(D_1[nT], D_2[nT])$$

where

$$D_1[nT] = \sum_{i=1}^{N} \frac{(E_i[(n+1)T] - E_i[(n-1)T])^2}{TE[nT]}$$

$$D_2[nT] = \sum_{i=1}^{N} \frac{(E_i[(n+2)T] - E_i[(n-2)T])^2}{TE[nT]}$$

$TE[nT]$ = the total energy across the entire spectrum at time nT

$E_i[nT]$ = the energy value at time nT in the ith energy bank

N = number of energy banks

T = 5 msec

Computed every 5 msec, S[n] has a high value at onsets and offsets and a low value in the steady state portion of vowels as shown by the arrows in the example of Figure 3-12. The reciprocal of the average value of spectral change is obtained from the middle portion of each syllable region so as not to take into account the spectral change produced by the onset or offset of the segments.

3.4.2 Combining the Features

Selected features may manifest stress information but it is the combination of the features that provides the mechanism for an automated stress recognition scheme. Several algorithms were investigated before an effective one was developed. The most significant predecessor of the current algorithm is a K-means clustering approach.

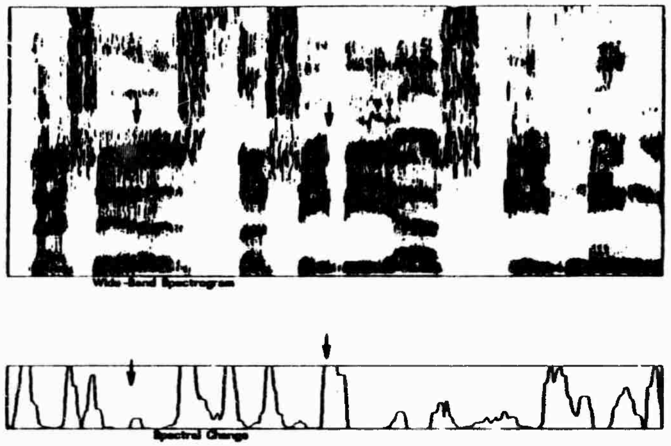


Figure 3-12: Spectral change measure is high for onsets and offsets and low in steady state portions of segments as shown by the arrows

A K-means clustering algorithm was implemented on a database of 100 isolated words or approximately 250 syllables. The normalized parameters computed on all the syllables provided the input data. The output of the clustering was two centroid values, a stressed and an unstressed centroid. Each syllable was associated with a parameter vector consisting of the five computed normalized parameter values. For each word, a Euclidean distance was measured from the centroids to each syllable parameter vector. The syllable with the minimum distance to the stressed centroid was designated as stressed. The remaining syllables were marked as unstressed. The results of this experiment were very discouraging with no better than 60% accuracy.

These results may improve as the amount of clustering data is increased. Performance may also be enhanced if there is some normalization for timing. In general, the sonorants in a four syllable word will be relatively shorter in duration than those of a two syllable word. If clustering were performed separately on two, three, and four syllable words, there may be some duration normalization that would improve its performance. However, this algorithm suffers from the assumption that there exists a global model for a stressed and an unstressed syllable such that any input token can be labeled as belonging to one of those bins. In actuality, inter- and intra-speaker variability corrupt the existence of such global models for phonemes. As an example, a stressed lax vowel in one word may be shorter in duration and lower in energy than an unstressed tense vowel in another word. In essence, the judgment of stress can be viewed as a relative one within a given word. Making judgments across words becomes dangerous as it is extremely difficult to normalize for the effect of phonetic and speaker variation.

In the current algorithm each input word is treated independently, unlike the clustering algorithm, so that the judgment of stress remains a relative decision. Each syllable is associated with a five dimensional feature vector, the components being the five normalized parameter values as demonstrated in Figure 3-13. In this figure, each sonorant region of a three syllable word is represented as S_i, where i ranges from 0 to 2. Each parameter is represented as the symbols A, B, C, D, and E. Each parameter value in each syllable is denoted as A_i, B_i, etc, where i denotes the appropriate syllable location.

In addition to the five dimensional feature vectors[3] associated with each syllable, a maximum feature vector is constructed from the maximum value of each parameter across the syllables, as shown below. In this example, A_0, B_1, C_0, D_2, and

[3] The notation for a vector is ~.

syllables parameters	S_0	S_1	S_2
A	A_0	A_1	A_2
B	B_0	B_1	B_2
C	C_0	C_1	C_2
D	D_0	D_1	D_2
E	E_0	E_1	E_2

Figure 3-13: Example of the representation of stress parameters for a three syllable word

E_0 correspond to the maximum values of parameters A, B, C, D, and E, respectively. The Euclidean distance from each syllable vector to the maximum vector is computed. The syllable with the minimum distance is designated as stressed with the remaining syllables marked as unstressed. A Euclidean distance is used instead of a distance weighted by covariance in order to keep the decision process a relative one within the word. A covariance matrix requires accumulating statistics over a large number of words similar to the clustering algorithm.

$$\tilde{S}_0 = (A_0\ B_0\ C_0\ D_0\ E_0)$$

$$\tilde{S}_1 = (A_1\ B_1\ C_1\ D_1\ E_1)$$

$$\tilde{S}_2 = (A_2\ B_2\ D_2\ C_2\ E_2)$$

$$\tilde{MAX} = (A_0\ B_1\ C_0\ D_2\ E_0)$$

In the cases where another syllable's distance is close to the minimum distance as determined by a threshold, two additional courses of action are taken. First, a likelihood measure based on the initial lexical studies is added. The likelihood is simply a number assigned to each syllable that is the probability of primary stress falling on that syllable depending only on the number of syllables in the word. As an example, only 39 of the 3000 four syllable words have stress falling on the last syllable. In contrast, 1343 of the words have stress on the second syllable. Thus, the second syllable has a likelihood of (1343 + 3000) or 0.45. Similarly, the fourth syllable has a likelihood of receiving stress of only 0.01. Thus, another parameter is added to each syllable that corresponds to the appropriate likelihood of receiving stress in that syllable. The feature vectors and distances are recomputed in a six dimensional space instead of a five dimensional one. Thus, the incorporation of lexical information acts as a tie breaker when the acoustic correlates are ambiguous.

The second course of action, for words with three syllables or more, is to provide a second choice for stress. In many cases, this may correspond to a secondary stress in the word such as the first syllable in the word *Massachusetts*.

After the stressed syllables are assigned, the duration and energy of the unstressed syllables are reexamined in order to make a reduced decision. If the duration is very short (less than 40 msec) or the duration is short (less than 60 msec) and the energy in the lower half of the spectrum (750 to 4000 Hz) is low, then the syllable is marked as reduced.

The current algorithm provides the following benefits. First, the method is a well defined algorithm that does not rely on the development of a set of rules that may only represent the performance of the examined database. Second, the algorithm is independent of the features and fairly modular. This establishes the system as a development tool for future research into the selection and evaluation of correlates for stress. Finally, the assignment of stress is based on a relative comparison of the syllables and does not rely on a global model of a stressed or unstressed syllable. This provides a timing normalization for differences in word lengths and number of syllables as well as inherent speaker timing differences.

To continue with the example discussed in the previous section, a spectrogram for the word *anxiety* is shown in Figure 3-14a. Directly below the spectrogram is the output from the front-end processor marking the syllable regions. The stress parameters are shown in Figure 3-14c-f where syllable average and maximum values are overlaid on the waveforms as appropriate. After the stress algorithm extracts the feature vectors and computes the distance metric, the final outpu of the system is shown in Figure 3-14g.

3.5 Lexical Access

In a particular application of isolated word, large vocabulary recognition, lexical stress provides strong constraints for lexical access. In this portion of the thesis, the lexical constraints provided by stress are manifested by a lexical access component. This component demonstrates the lexical properties investigated in the early part of

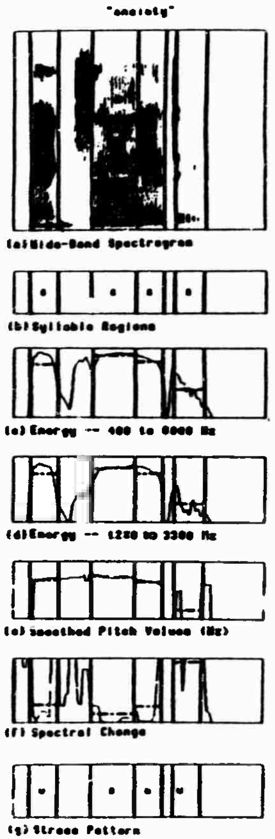


Figure 3-14: Example of the work done by the stress algorithm for the word *anxiety*

this thesis. In addition, the incorporation of lexical information completes the system as a development tool for further research into stress and its lexical constraints.

The lexicon is based on the polysyllabic words of MPD. The orthography for each word is associated with one pronunciation baseform which includes a marker for primary stress. Syllables are determined automatically by a search for vowel or syllabic segments in the phonetic transcription. These words, approximately 16000, are mapped into stress pattern equivalence classes according to their dictionary phonetic transcriptions.

The baseforms are also expanded by a set of phonological rules that account for segment reduction, transformation, or deletion. Rules governing the phonological behavior of phonetic segments have been established by Oshika, et al [21]. For example, a vowel deletion rule states that a vowel between two consonant clusters following a syllable of greater stress and preceding a syllable of equal or greater stress can be deleted as in *boundary* or *interest*. Further constraints must be imposed on the consonant clusters surrounding the deleted vowel. The preceding consonant must be an allowable syllable final cluster in English and similarly the following consonant cluster must be an allowable syllable initial cluster. Allowable syllable final clusters are limited to single consonants and a few clusters such as /st/, /sp/, and /ks/. Allowable syllable initial clusters are far less constrained. Additional rules predict the deletion or devoicing of vowel segments.

The importance of these rules for speech recognition is that the number of syllables in a word may be transformed acoustically and perceptually. The lexicon can be expanded by these phonological rules and then classified again into stress pattern classes. (See Appendix B for the rules incorporated.) As an example, the word *chocolate* has a dictionary representation of /čɔkʌlɪt/ but may be expanded by rules into the following pronunciations: /čɔkəlɪt/, /čɔklɪt/, and /čɔklɪt/. Thus, the expansions of the word *chocolate* would correspond to the equivalence classes

[SUU], [SRU], [SU], and [SR], respectively. There are approximately 100 equivalence classes in all and a total ex, nded lexicon of 25000 words.

The output of the stress algorithm is a time-aligned stress pattern for each input word. From the derived stress pattern, lexical access provides a list of word candidates. There could be at least three reasons why the input would not be a member of the candidate list. First, if the front-end processor distorted the number of syllables such that even the phonological rules could not produce such a pronunciation, then an error would result. Second, a mislabeling of the primary stress is an unrecoverable error for lexical access. Finally, if the actual pronunciation of the word can not be accounted for by the phonological rule expansion, then the input token will not be a member of the candidate list.

Again using the example of the word *anxiety*, lexical access produces a set of word candidates associated with the derived stress pattern, [USUU]. There are 1300 words that fall into that equivalence class with *anxiety* being one of them. This equivalence class is 5% of the total expanded lexicon which is less than the expected class size of 15% derived in Chapter 2.

3.6 Chapter Summary

In this chapter, we discussed the following.

- The benefit of stress information for speech recognition is based on its ability to illuminate areas of acoustic reliability. In addition, for the specific recognition scheme of lexical access, stress information provides strong lexical constraints.

- The implementation of a stress recognition system has three major components: a front-end processor to derive syllable regions; a stress algorithm that extracts feature vectors for the syllables and makes a judgment of stress based on their relative merits; and finally a lexical access component that maps the derived stress pattern into a large lexicon to yield a list of word candidates.

The front-end processor and the stress algorithm provide a prosodic analyzer that would benefit large vocabulary isolated word recognition systems such as the one built by Huttenlocher [10]. The lexical access, on the other hand, serves mainly to complete this prototype stress recognition system and to demonstrate the lexical properties of stress as discussed earlier in this thesis. Lexical representation and the derivation of phonological rules to account for various acoustic realizations were not the primary focus of this thesis.

Chapter Four

Performance Evaluation

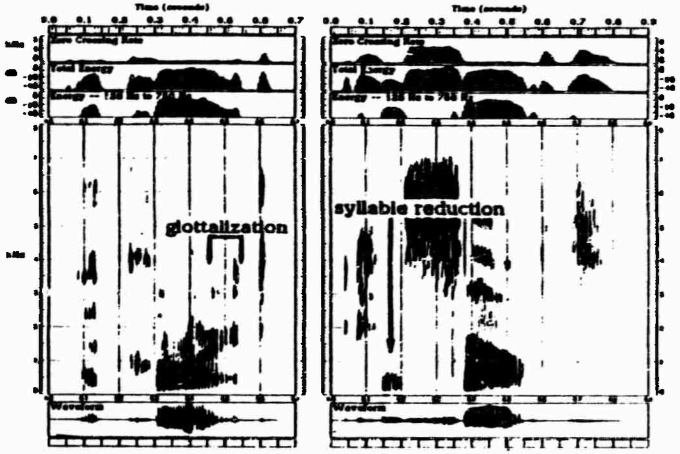
This chapter describes the evaluation criteria for the system and its subsequent performance. For this isolated word system, there are essentially two classes of error, segment and word error. Segment[4] error examines the data on a segment by segment basis. Word error is more strict in that the entire word must be free from error to be correct. As an example, if the evaluation data consisted of 2 words, two syllables each, then there would be 2 words and 4 segments to evaluate. If one of the syllables was not found by the system, then the segment error rate would be 25% and the word error rate would be 50%. As another example, since there are two stress segments in two words, a mislabeling of one of the stressed syllables by the stress algorithm would yield a segment and word error rate of 50%. Clearly, word error presents a more strict means of evaluation than segment error and will generally be greater than or equal to segment error.

The evaluation is broken down into the three components of the system: the front-end processor; the stress algorithm; and lexical access. The stress algorithm is the most important component of the system as it establishes islands of acoustic reliability independent of any specific recognition scheme. The front-end processor determines the sonorant regions of the syllables which are passed to the stress algorithm. Locating 100% of all the syllable units is not necessary for the stress algorithm to find the most stressed region. In other words, the location of the stressed region is valuable for acoustic reliability even when the number of syllables is not known exactly.

[4] In this evaluation, *segment* corresponds to a syllable unit.

As discussed earlier in this thesis, stress information can also be valuable in the particular recognition scheme of large vocabulary, isolated word recognition based on lexical access. Establishing the position of stress and the number of syllables in a word provides strong constraints on the word candidates proposed by lexical access. The lexical access component of the system is an attempt to demonstrate this principle. It relies heavily on deriving the correct number of syllables in the word as well as the position of stress. This is a difficult task especially in isolated words. The word ending effects such as glottalization or extremely reduced syllables, shown in Figure 4-1, make syllable identification difficult as these segments no longer carry strong sonorant characteristics. Phonological rules can account for some of the variation, such as the reduced syllable deletion in the word "interest". However, there is a great deal of phonological variation that either does not have an appropriate set of rules written or is speaker dependent such that general rules cannot compensate. Thus, in order for the lexical component to be highly successful, better phonological rules need to be investigated to account for observed acoustic realizations. In addition, the syllabification process needs to be extremely accurate since deleting or inserting a segment can be a major error for lexical access. These issues are not the focus of this thesis and will not be addressed extensively in this evaluation. If the stress is mislabeled, the lexical access can not recover from such an error. The evaluation of the lexical component will only be on data that is free from front-end and stress errors.

The emphasis of the evaluation is on the stress algorithm and how well it labels stressed segments independent of how the syllable information is obtained. Since the system is designed to be practical for speech recognition, the syllable information is extracted directly from the speech signal and does not rely on a transcription being provided. Thus, we must evaluate the performance in this realistic environment in order to demonstrate the usefulness as well as deficiencies of such a system.



"accord"
(a)

"consult"
(b)

Figure 4-1: Wioe-band spectrograms showing word beginning and ending effects such as (a) glottalization and (b) extreme syllable reduction

4.1 Evaluation Data

The evaluation data is composed of 1600 isolated words (4500 syllable segments), spoken by 6 male and 5 female speakers, and digitized at 16 kHz. The word corpus distribution, shown in Figure 4-2, indicates that 80% of the words are two and three syllables while the remainder are four and five syllable with varying stress patterns and phonetic contexts. The distribution is similar to that of MPD to ensure that the evaluation data is representative of possible input data. In order to guarantee an objective evaluation, the corpus of words is completely different from that used in the analysis and system development.

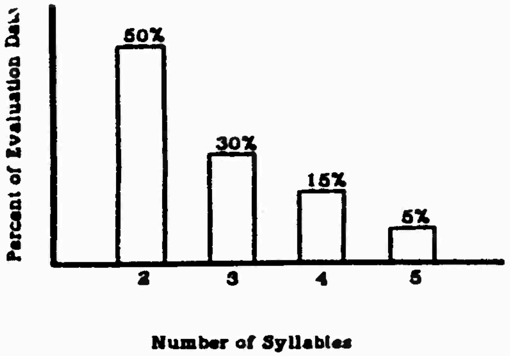


Figure 4-2: Distribution of evaluation data by number of syllables

As discussed in previous chapters, there are different degrees of difficulty involved in extracting the sonorant regions corresponding to each syllable. In the word *Massachusetts*, each sonorant region corresponds to a syllable which somewhat simplifies the front-end processing task. In contrast, the word *yellow* is difficult to separate into syllable regions due to the intervocalic sonorant, /l/. A word such as *anxiety* is even more difficult since there is no consonant across the syllable

boundary, only the vowel sequence, /Y E/. As the difficulty of the data increases, the system performance should deteriorate. In order to distinguish between the performance of the front-end processor and the inherent difficulty of some data, the evaluation data is subdivided according to the level of difficulty of syllable separation. (See Appendix C for words in the various data groups.) The division of the data in increasing order of complexity is as follows:

1. Group 1: easy data --- syllables separated by obstruents such that no further segmentation is necessary after initial broad segmentation;

2. Group 2: nasals and semivowels --- syllables separated by intervocalic nasals and the liquids, /l w/;

3. Group 3: vowel --- no consonant between syllables, only a sequence vowels;

4. Group 4: /r/ --- syllables separated by intervocalic /r/.

The distribution of the MPD by the above classes is shown in Figure 4-3a. Twenty-five percent of the words have at least one occurrence of an intervocalic nasal or liquid. Intervocalic /r/'s occur in 11% of the words while 9% have a sequence of two vowels. Sixty-three percent of the words have at least one obstruent separating the syllables. It is encouraging that the majority of words fall into this last category as that simplifies the amount of work that needs to be done in the front-end processor of the system.

The distribution of the evaluation data by intervocalic context is similar to that of MPD as shown in Figure 4-3b. In the data, 74% of the words have obstruents separating syllable boundaries while 26% of the words fall into more difficult categories.

In addition to the data categories above, there are 250 words in the database that form stress word pairs such as *CONtrast/conTRAST*. There are 50 words of this

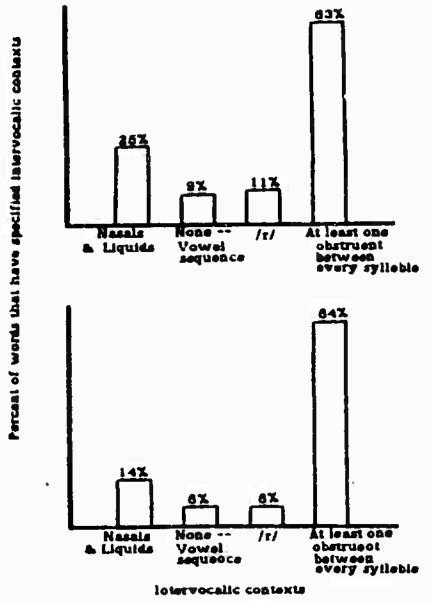


Figure 4-3: Distribution by classes of intervocalic contexts of (a) MPD and (b) the evaluation data

nature. (See Appendix D.) They were read in isolation by five speakers from a list that had the appropriate stress marked. Two of the five speakers were used in the previous data as well. These words, all but one pair containing two syllables, are typically the most confusable from a segmental representation, and their stress patterns should play a major role in distinguishing them. In addition, since the speakers were told explicitly where to stress each of the words, this data may provide

an upper bound on performance. It is possible that the emphasis would be more robust in these cases than when the stress is left open to the interpretation of the speaker.

The evaluation of the system as a whole relies on the above breakdown of the evaluation data. Specifically, the performance of the front-end processor is readily captured only when the difficulty level of its task is known.

The appropriate stress pattern for each word in the database is usually obvious from a visual examination of a spectrogram, and, in these cases, it is relatively easy to establish the stress pattern. For all the cases in which the stress algorithm misidentifies the location of stress, listening tests are performed similar to those described in Chapter 2. Five listeners listen to a tape of the words missed by the system with each word repeated twice. They are asked to mark the position of greatest stress or mark as ambiguous if they can not perceive the location of stress. For these cases, if the majority of listeners do not agree on the location of stress, then there is some inherent ambiguity and the system error is more acceptable. Algorithm errors can be penalized less if reinforced by human judgment.

The next sections of this chapter discuss the evaluation and performance of the individual components of the system.

4.2 Evaluation of the Front-End Processor

The front-end processor is composed of two stages. First, the broad phonetic segmentation [17] identifies regions that are sonorant. Sources of error in the segmentation consist of labeling a sonorant region as a member of another phonetic class, or labeling a non-sonorant as sonorant. In other words, segment deletion or segment insertion are the prime sources of error.

The second stage of the front-end processor examines the sonorant regions derived from the initial segmentation for additional syllable boundaries. The errors that could stem from this further segmentation are identifying a syllable boundary when there is none present or not labeling a robust syllable boundary. Again, segment insertion and deletion are the causes of error in this further segmentation stage.

The performance of the front-end processor, including the initial and further stages, is shown in Table 4-1. Each column represents one of the evaluation data groups with the last column summarizing the entire database. In this evaluation, the word pairs are included with the first group of data as they are relatively easy to segment. The three main rows correspond to the evaluation of the initial segmentation, the further segmentation, and finally the front-end processor as a whole. Each main row is then divided into segment error and word error as discussed previously.

In the total database there are 1600 words and 4500 segments (or syllables). The initial segmentation makes an error on 3% of the segments and on 7% of the words. The further segmentation, however, has an error in only 1% of the segments and 2% of the words. The front-end processor as a whole, run on the total evaluation database, makes 4% and 9% segment and word error, respectively.

A point of interest is that the performance of the further segmentation becomes increasingly worse as one progresses from group 1 to group 4 of the data, especially in terms of word error. These results seem appropriate since the job of this stage of the front-end processor is getting harder in each data group. The initial segmentation, however, is unaffected by the level of difficulty of the data groups with the exception of Group 3. In this group, it is more common for a sonorant to fall at the end of the word. Glottalization is thus more prevalent affecting the performance of the initial segmentation. Otherwise, the location of sonorant regions by the initial segmentation has a steady level of difficulty across the data.

Segmentation Error \ Data Group		1	2	3	4	Total Data
Initial	segment	3%	2%	4%	1%	3%
	word	7%	7%	12%	3%	7%
Further	segment	1%	1%	2%	2%	1%
	word	2%	4%	6%	8%	2%
Front-end Processor	segment	4%	3%	6%	3%	4%
	word	9%	9%	18%	11%	9%

Table 4-1: Evaluation of front-end processor by data groups and by segment versus word error

In order to understand the types of error in the front-end processor, the evaluation can be broken down into segment deletion and insertion errors, shown in Table 4-2. Similar to the previous table, each column represents a data group with the last column containing the entire database. The three rows show initial segmentation, further segmentation, and total front-end processor error breakdowns. In addition, each row is divided into deletion errors and insertion errors. All the errors in this table deal with segment error as opposed to word error.

Segmentation Error		1	2	3	4	Total Data
Initial	deletion	.8%	.3%	1.9%	0%	.8%
	insertion	2.4%	1.7%	2.5%	.9%	2.2%
Further	deletion	0%	.7%	1.6%	2.1%	.4%
	insertion	.8%	.3%	0%	0%	.6%
Front-end Processor	deletion	.8%	1%	3.5%	2.1%	1.2%
	insertion	3.2%	2%	2.5%	.9%	2.8%

Table 4-2: Evaluation of front-end processor by data groups and by segment deletion versus segment insertion error

Looking at the first column of Table 4-2, the easy data group, one can see that 0.8% of the errors is due to deletion of segments by the initial segmentation. The further segmentation stage deletes no segments in this data. The insertion of extraneous sonorant segments contributes to 3.2% of the error, 2.4% from the initial segmentation and 0.8% from the next stage. The amount of deletion by the further segmentation stage increases as the degree of difficulty of the data increases. In

other words, as more difficult segments (such as intervocalic /r/'s) become present, the detection of them deteriorates as was originally presumed. The insertion error rate of this stage decreases, however, as there is less opportunity for false alarms due to the phonetic contexts.

The major contribution of error to the front-end processor is from the initial segmentation. Its 7% total word error rate, as compared to 2% for the further segmentation, can be partially attributed to the difficulty in labeling the beginning and ending portions of isolated words. Sixty percent of the missed segments are unstressed and either at the word beginning or word ending. In fact, all of the initial segmentation errors are in unstressed syllables where the acoustic information may be highly variable.

In total, 2.8% of the error is due to segment insertion while 1.2% is due to segment deletion. In addition, 3% of the error, deletion and insertion, is from the initial segmentation. The further segmentation contributes only one fourth of the front-end processor segment error. All of the insertion errors of the further segmentation are in stressed syllables where formant motion is more pronounced and thus more likely to be classified as two segments instead of one. All of the deletion errors in this stage are in a post stressed position which receives less emphasis than a prestressed position. Thus, these segments are less robust acoustically and hence harder to detect.

In summary, the 4% total segment error rate for all the data may be adequate as a front-end to a stressed segment detection system since stressed segments can be correctly identified even if segments are missing or are inserted. However, if the entire stress pattern for each word is to be derived, a harsher demand, the 9% word error rate for the front-end processor significantly decreases the ability to do so, as determining the stress pattern demands correct identification of all the syllable segments.

4.3 Evaluation of the Stress Algorithm

In the evaluation of the stress algorithm, there are two main perspectives, in increasing severity, from which one can view the results. The first is how many of the words in the database have the stressed region correctly labeled. In other words, despite the front-end processor errors, the location of stress is derived. This is important as it provides pointers to regions of acoustic reliability which can benefit almost any phonetically-based recognition scheme. It is also the most practical to achieve. Second, if the entire stress pattern is correct, a lexical mapping can be used to propose word candidates. This is the most strict means of evaluation as there is no room for front-end processor or stress errors.

The first means of evaluation concerns the labeling of the stressed segments. Looking at the total corpus of evaluation data, 98% of the stressed segments are identified as such. In other words, out of 1600 words, 1568 of them have the stressed syllable correctly labeled. For the remaining 2% (32 words) that are mislabeled, the following is a breakdown of the causes of error.

- Nearly 40% of the labeling error is due to front-end processor errors: 25% from segment insertion; 15% from segment deletion. Such errors influence the parameters measured and thus the judgment of stress.

- In 30% of the labeling errors, the stressed syllable is marked as a second choice for stress in words with three or more syllables. In these cases, the acoustic features of the syllable marked as stressed and as alternate stress are similar. The second choice is a reasonable course of action because it says there may be two regions in a word that are acoustically reliable. Such is often the case with primary and secondary stress.

- In 8% of the mislabeled stress segments, the error is due to a confusion of the stressed segment with the word final segment. In these cases, prepausal lengthening is unusually long and thus under compensated.

- The remaining 22% of the mislabeled stress segments are more difficult to justify. However, listening tests reveal that two thirds of these words are ambiguous to the listeners. In other words, the majority of listeners do not agree on the location of stress for two thirds of the remaining 22% with labeling errors.

For a moment, let us isolate the performance of the word pairs data. The data consists of minimal stress pairs such as *CONtrast/conTRAST*. The only three syllable word is the pair *ATtribute/atTRIbute*. All of the rest are composed of two syllables. In this group of data, only 1.3% (3 words) of the stressed segments are labeled as unstressed. In addition, there are no pairs that suffer from a mislabeling in both words of the pairs. The system performs best on this data group, as anticipated, since the speakers were told explicitly where to put the stress in each word.

An interesting aspect of stress labeling error may be the breakdown of the data by number of syllables. In Figure 4-4a, the histogram shows the distribution of the evaluation data by number of syllables in the words. Figure 4-4b depicts the percent of stress segment labeling error in 2, 3, 4, and 5 syllable words. The data and the error distributions are similar. For example, two syllable words, which comprise 50% of the corpus, are subject to 50% of the labeling error. Thus, the labeling error does not directly depend on the number of syllables in the word.

An interesting aspect of evaluation concerns the ability of the system to derive the correct syllable locations as well as the stressed segment location. This is a more strict means of evaluation than just locating the stressed segments. In the entire corpus, 90% of the words meet this requirement, a slightly better result than the previous evaluation. Comparing with the results on stressed segment identification presented previously, we see that of all the words that have the correct location of stress, 92% of them also have the syllables correctly identified.

The second evaluation of the stress component assesses the entire stress pattern for the word. This means that there can be no front-end processor error, no stress labeling error, and no confusions of unstressed and reduced (schwa) segments. This is harsh but necessary if it is to be used as the input to lexical access.

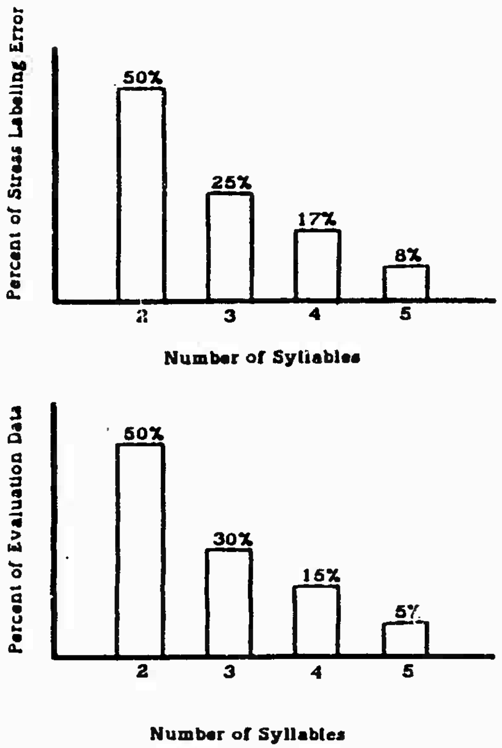


Figure 4-4: Distribution of (a) number of syllables in database and (b) number of syllables in mislabeled stress data

For the entire database, the stress pattern is correct in 87% of the words. This implies that in 3% of the words, there is confusion between reduced and unstressed segments. The confusion is evenly distributed in that nearly half of the confusions are unstressed segments labeled as reduced, where the other half of the error labels reduced segments as unstressed. Sixty percent of these confusions are due to the boundaries established by the initial segmentation. When boundaries are too wide or narrow for the actual segment, the duration measured between the boundaries is inaccurate and may cause an incorrect labeling. The remaining portion of unstressed/reduced confusion error may be attributed to the fact that the reduced decision is based on duration and energy thresholds which may be exceeded in some extreme instances.

In summary, the system performs well in stress identification. Its performance deteriorates as more demands are placed on the identification of all the syllable units. Therefore, this system is strong for determining areas of acoustic reliability and has less strength in the application of lexical access where the entire stress pattern is necessary.

4.4 Evaluation of Lexical Access

The performance of the lexical access component is dependent on the performance of previous stages. If the correct number of syllables are not found or the stress is mislabeled, then the lexical access will fail. Furthermore, if a syllable is marked as reduced by the stress component that can not be reduced in the lexical representation by the phonological rules, then again lexical access will fail. Thus, this component is very sensitive to propagation of errors. At the same time, this component of the system is designed to merely demonstrate some of the lexical constraints investigated in the lexical study of Chapter 2. It is not intended to be a functional system in itself. In fact, the optimal application of stress information in such a specific recognition task is in conjunction with segmental information. It is the isolated stress information that proves most useful in a recognition such as that of Huttenlocher discussed previously.

A successful lexical access not only needs accurate input but it also must have a thorough lexical representation. The phonological rules that expand the baseforms should capture any acoustic realizations that may occur as input. Unfortunately, the writing of appropriate phonological rules is an immense project in itself and is not the emphasis of this thesis. Instead, a more general yet well established set of phonological rules is incorporated just to provide a test bed for lexical access.

The evaluation of lexical access is based on only the 87% of the words in the evaluation corpus that are free from front-end processor and stress errors. Successful word retrieval (that is, the input word is a member of the candidate list proposed by lexical access) occurs in 96% of the data. The errors are due to acoustic realizations of words that are not represented in the expanded lexicon. For example, the word *conceal* has a lexical representation of /kʌnsi'l/. However, one acoustic realization of it has three syllables, /kʌnsi'ɨl/, for which there is not a phonological rule to produce it from the baseform.

4.5 Chapter Summary

The following are the major points of this chapter.

- The system is evaluated on a database of 1600 isolated words with varying phonetic contexts and stress patterns. The evaluation data is completely separate from the data used in the system design and analysis.

- Given an isolated word with approximate syllable identification, the stress system correctly labels the region of stress in 98% of the words.

- The system is less strong in deriving accurate syllable identification than it is in labeling the acoustically robust regions. These demands are harsher and thus the performance suffers. The number of syllables and the location of stress are correct in 90% of the words.

- The determination of the entire stress pattern demands correct syllable identification as well as no confusion between unstressed and reduced segments. This is the most strict stipulation and the system performance, appropriately, drops to 87% of the words being labeled with the correct stress pattern.

- Lexical access, based on stress pattern information alone, is difficult as it demands nearly perfect stress and syllable labeling. In addition, it is not clear that it is meaningful since a recognition framework is benefited most by the combination of prosodic and segmental information. The lexical access, contributing an additional 4% to the system error, also suffers from a lack of the translation of acoustic realizations into well-formed phonological rules.

Chapter Five

Summary

In this thesis, we have addressed the issue of lexical stress determination. Our motivation is twofold. First, stress provides pointers to areas rich in acoustic information in the speech signal. This is important for phonetically based recognition schemes as it establishes regions of reliability where more detailed information may be extracted with greater confidence. Second, in a particular recognition scheme based on lexical access, stress information, as well as the number of syllables, provides strong constraints.

The well-known acoustic correlates of stress, namely duration, energy, and fundamental frequency, are combined in the implementation of a stress recognition system. In the designed system, initial segmentation derives the sonorant regions corresponding to each of the syllables within an isolated word. Each syllable is then associated with a feature vector consisting of the correlates measured. A relative comparison of the feature vectors within the word establishes the location of stress. The lexical component then maps the derived stress pattern into a large expanded lexicon to yield a set of word candidates.

The strength of the system lies in its ability to identify regions of stress which correspond to islands of acoustic-phonetic reliability. The stress pattern is more difficult to achieve as it relies on finding the correct number of syllables as well as making a correct distinction between unstressed and reduced syllables. The performance of the system deteriorates as the difficulty in locating syllables increases. The stress algorithm is useful in itself, independent of any recognition scheme. The lexical component, however, merely demonstrates some of the lexical properties presented earlier.

5.1 Suggestions for Further Research

The evaluation sheds light on some aspects of this work that would be interesting for further research or improvement. In particular, there are four areas that may be worth pursuing.

First, the initial broad segmentation used may not be the most appropriate front-end. This segmentation [17] is designed as an initial step in a automatic alignment system that aligns the speech signal with a user-provided phonetic transcription. Labeling errors in the segmentation are accounted for in the alignment of the segmentation labels with the phonetic transcription labels through a set of cost functions. Thus, the emphasis of the segmentation is on boundaries as opposed to labels. Of course, the segmentation algorithms in many systems suffer from a lack of robust labeling. It is a difficult problem in speech and the solution is not entirely clear at this point. Work done on syllable extraction, such as that by Mermelstein [19], may present a more appropriate classification of the acoustic signal since detailed labeling is not necessary for the stress recognition. Syllable recognition may also be more robust.

A second area of interest for further work is in the further segmentation where additional syllables are located within sonorant regions. The spectral weighting used to illuminate formant changes is good for proposing possible transitions. However, in order to verify that the transition is actually two vowels or an intervocalic semivowel, further features should be extracted. In other words, specific features depending on the type of boundary proposed could prevent diphthongs, syllable-final sonorants, or just extremely long vowels from being characterized as two syllable units.

Third, if the lexical component is to be valuable in a particular application, its success depends on accurate syllable extraction and a thorough set of phonological rules by which to expand the lexicon. It is necessary to achieve a more elaborate understanding of the relationship between the acoustic realizations and the phonological rules that emulate them. There is a difficult trade-off between writing rules that create all the variations observed and writing a set of coherent, general, and justifiable rules. It is certainly a difficult problem that relies on expertise in acoustic-phonetics and phonology.

A pathfinding algorithm may be a suitable alternate approach to the lexical mapping discussed previously. The stress pattern derived by the system could be aligned with the lexical representations. Training data could be used to formulate a set of cost functions used by the pathfinder. The strength of this method, in general, is that errors, such as segment deletion, insertion, or mislabeling, would be handled by the alignment algorithm and the appropriate cost functions. The lexical component would rely less on previous stages of the system or on lexical expansions.

Finally, the fourth area concerns the nature of the decision made by the stress algorithm. Currently, the algorithm forces a segment to be labeled as stressed, unstressed, or reduced. An alternative is to soften the decision by assigning a score to each segment corresponding to "goodness" of the stress features measured. A scoring scheme may better illustrate the confidence leval associated with each assigned label. In addition, the relative scores of the segments within a word would be clearly manifested.

In addition to further work for improvcment, one must also consider the extensibility of a system. The most important aspect of this is the extension of the stress algorithm from isolated words to continuous speech. Currently, the algorithm makes a stress decision by means of a relative comparison of all the syllable feature vectors within a word. In a continuous sentence of speech, there may be several points of stress. Instead of picking the most stressed candidate, the algorithm would have to choose the best *n* regions of stress. A technique such as clustering may be

appropriate for sorting the data points, or syllable units, into bins of stress and unstressed. The current algorithm provides a good means of feature extraction, but for continuous speech the comparison of the feature vectors and the decision process needs to be more generalized.

Appendix A

Word corpus

This is the word corpus used in the initial acoustic study discussed in Chapter 2.

LAZY
MEASURE
HAMMER
COMIC
TECHNIQUE
UTMOST
KITCHEN
MOISTURE
OMNIBUS
SOMEBODY
DESCRIPTION
INTERNATIONAL
ANIMAL
ASSOCIATE
SOLUTION
MEDICINE

DENIES
DEMISE
ETHNIC
MAJOR
RELIES
ADMIT
HEAVY
JUSTICE
HESITATE
DAMNATION
PROFESSOR
ENIGMA
WASHINGTON
TRIUMPHANT
AMBASSADOR
MASSACHUSETTS

HELPMATE
FLAVOR
DEVISE
WEAKNESS
DIRTY
YELLOW
FORGIVE
INSULT
ABNEGATE
HOSPITAL
DALMATION
VOLTMETER
EXAMPLE
INSTITUTE
CELEBRATE

PICNIC
COGNATE
FINGER
CONIC
CREATURE
LEMON
ANGER
WORSHIP
YESTERDAY
SATISFY
OFFICIAL
NEWSPAPER
INVESTIGATION
PACIFIC
DECIMAL

Appendix B

Phonological Rules

The following are the phonological rules used to expand the lexicon from its baseforms. Their intent is to allow for changes in the number of syllables in the actual pronunciation from the baseforms

Vowel deletion rule:

$$[\text{V}, \beta \text{ stress}] \dashrightarrow \emptyset \;/\; [\text{C}_0]\,[\text{V}, \alpha \text{ stress}]\,[\text{C}_1] \text{ -- } [\text{C}_1]\,[\text{V}, \gamma \text{ stress}]$$

where $\text{V} \equiv$ vowel

$\text{C}_1 \equiv$ one or more consonants

$\text{C}_0 \equiv$ zero or more consonant.

$\alpha > \gamma >= \beta$

and C1 is an allowable syllable final or initial cluster, respectively

Schwa devoicing rules:

$$[\text{schwa}] \text{ ---> } \emptyset \;/\; [\text{voiceless stop}] \text{ -- } [\text{voiceless stop}]$$

$$[\text{schwa}] \text{ ---> } \emptyset \;/\; \# \,\text{C}_1 \text{ -- } \text{C}_1 \,[\text{stressed vowel}]$$

Vowel reduction rule:

unstressed short vowels ["I" "E" "Λ" "U"] can be reduced

Diphthong-glide rule:

```
["r" or "l"] ---> ["R` or "L"] / [DIPHTHONG] -- #
     where DIPHTHONG ∈ ["e" "W" "Y" "O" "u"]
```

Appendix C

Evaluation Data

This is the evaluation database divided into the levels of difficulty of syllable extraction as discussed in Chapter 4.

C.1 Evaluation Data: Group 1

ABSORB
ACCEPT
ACCORD
ACHIEVE
APART
APPEND
APPOINT
ASCEND
ASHAME
ASHORE
ASSAULT
ASSEMBLE
ASSEMBLY
ASSERT
ASSIGN
ASSIST
ASSISTANCE
ASSUME
ATTACH
ATTEND
BECAUSE
CAPACITY
COMPOSE
CONCEAL
CONCEIVE
CONCERN
CONDUCT
CONSENT
CONSIDER
CONSIST
CONSULT
CONSUME
CONTENT
DECEIVE
DECIDE
DECISION
DEPART
DEPEND
DEPOSIT
DESCEND
DESERT
DESERVE
DESIGN
DESIRE
DESPITE
DIGEST
DISASTER
DISEASE
DISGUISE
DISTURB
ENTITLE
ESCAPE
ESSENTIAL
ESTATE
EXACT
EXAMPLE
EXCEED
EXCEPT
EXCITE
EXECUTIVE
EXHAUST
EXHIBIT
EXIST
EXPECT
EXPOSE
EXTEND
EXTENT
HOTEL
IMPOSE
IMPOSSIBLE
INSIST
INSTEAD
INSULT
INTEND
INTENT
MACHINE
MANKIND
MISTAKE
NECESSITY
OBJECT
OBSERVE
OBSERVER
OCCASION
OPPOSE
PACIFIC
PERCEIVE
PERCENT
PHYSICIAN
POSITION
POSSESS
POSSESSION
POTATO
PRESENT
PRESERVE
PRETEND
PROCEEDING

PROCEDURE	PROCESSION	PRODUCE	PRODUCTIVE
PROPOSAL	PROPOSE	PROTECT	PROTECTION
PURSUIT	RECEIPT	RECEIVE	RECORD
RECOVER	REDUCTION	REJOICE	REPORT
RESEMBLE	RESERVE	RESIGN	RESIST
RESORT	RESULT	RESUME	SENSATION
SINCERE	SUCCEED	SUCCESS	SUCCESSFUL
SUCCESSION	SUGGEST	SUGGESTION	SUPPOSE
TRANSPORT	UNCERTAIN	UPSET	VACATION
ANGER	WORSHIP	FORGIVE	FINGER
FLAVOR	MAJOR	MEASURE	CREATURE
WEAKNESS	DIRTY	EXTRA	HEAVY
INSULT	JUSTICE	KITCHEN	LAZY
MOISTURE	AMBASSADOR	DESCRIPTION	INVESTIGATION
HOSPITAL	WASHINGTON	YESTERDAY	MEDICINE
SATISFY	EXAMPLE	HESITATE	UTMOST
INSTITUTE	NEWSPAPER	OFFICIAL	PACIFIC
PROFESSOR	ABNEGATE	ADMIT	ARSENIC
BENDING	BOTTLE	BOTTOM	BUTTON
CAMPBELL	CAMPER	CHIPMUNK	COGNATE
CRITTER	ETHNIC	FONDEST	GISMO
HADDOCK	HANGER	HANGMAN	HAVOCK
HELPMATE	HUNGER	IGNORE	MATTER
OATEN	PANDER	PANTER	PARSNIP
PICNIC	PIGNUT	PINCHING	SENTRY
SINGER	SINKING	SKIMPY	SUNKEN
TECHNIQUE	TOTAL	MASSACHUSETTS	

C.2 Evaluation Data: Group 2

ECONOMY	ATTORNEY	PARTICULAR	ADDITIONAL
ETERNAL	DETERMINE	OCCASIONAL	INTELLIGENCE
MACHINERY	FACILITY	EXAMINE	MECHANICAL
INTELLIGENT	YELLOW	HAMMER	LEMON
CELEBRATE	SOLUTION	INTERNATIONAL	ANIMAL
BANNER	BANNOCK	CHIMNEY	COMIC

CONIC	DALMATION	DECIMAL	DEMISE
DENIES	ENIGMA	FLANNEL	HAMMER
HAMMOCK	HOMELY	INMATE	LONELY
OMNIBUS	POLISH	RELIES	RUNNY
SIMMER	SINNER	SULLY	TILLER

C.3 Evaluation Data: Group 3

MATERIAL	ANXIETY	INSURANCE
ENCOURAGE	SOCIETY	EXPERIENCE
AssOCIATE	INTERIOR	MACHINERY
SUPERIOR	MUSEUM	ASSOCIATE
TRIUMPH	TRIUMPHANT	

C.4 Evaluation Data: Group 4

MATERIAL	COMPARISON
INSURANCE	ENCOURAGE
EXPERIENCE	INTERIOR
MACHINERY	EXPERIMENT
MAJORITY	APPARENT
COMPARATIVE	SUPERIOR
DESIRABLE	

Appendix D

Evaluation Data: Stress worc pairs

These are the word pairs with contrasting stress used in the evaluatio1. The syllable that is capitalized denotes the stressed syllable.

EXport	exPORT
PERfect	perFECT
COMpress	comPRESS
INcline	inCLINE
CONtract	conTRACT
UPset	upSET
DIgest	diGEST
EScort	esCORT
COMpact	comPACT
COMpound	comPOUND
IMport	imPORT
REcord	reCORD
ATtribute	atTRIBute
REbel	reBEL
CONvert	conVERT
CONtrast	conTRAST
TRANsport	tranSPORT
INsult	inSULT
OBject	obJECT
CONduct	conDUCT
UPlift	upLIFT
CONtest	conTEST
TORment	torMENT
CONflict	conFLICT
SURvey	surVEY

References

[1] Bolinger, D. L.
A Theory of Pitch Accent in English.
Word 14:109-149, 1958.

[2] Chen, F. C.
A Continuous Digit Recognition System Based on Acoustic-Phonetic Knowledge.
Forthcoming Doctoral Dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

[3] Chomsky, N. and Halle, N.
The Sound Pattern of English.
Harper and Row, New York, Evanston, and London, 1968.

[4] Cole, R. A., Zue, V. W., and Reddy, D. R.
Speech as Patterns on Paper.
In R. A. Cole, editor, *Perception and Production of Fluent Speech*, pages 3-50. Lawrence Erlbaum Assoc., Hillsdale, N. J., 1980.

[5] Cole, R. A. and Jakimik, J.
How are Syllables used to Recognize Words?.
Journal of the Acoustical Society of America 67, March, 1980.

[6] Culter, A. and Foss, D. J.
On the Role of Sentence Stress in Sentence Processing.
Language and Speech 20, 1-10, August, 1977.

[7] Fry, D. B.
Duration and Intensity as Physical Correlates of Linguistic Stress.
Journal of the Acoustical Society of Amarica 27, No. 4. July, 1955.

[8] Fry, D. B.
Experiments in the Perception of Stress.
Language and Speech 1:126-152, 1958.

[9] Huttenlocher, D. P. and Zue, V. W.
Phonotactic and Lexical Constraints in Speech Recognition.
AAAI, August, 1983.

[10] Huttenlocher, D. P.
Acoustic-Phonetic and Lexical Constraints in Word Recognition: Lexical Access Using Partial Information.
Master's thesis, Massachusetts Institute of Technology, 1984.

[11] Klatt, D. H.
Vowel Lengthening is Syntactically Determined in a Connected Discourse.
Journal of Phonetics 3:129-140, 1975.

[12] Klatt, D. H.
Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence.
J. Acoust. Soc. Am. 59,1208-1221, 1976.

[13] Klatt, D. H.
Review of the ARPA Speech Understanding Project.
J. Acoust. Soc. Am. 62, No. 6:1345-1366, 1977.

[14] Laface, P.
A Format Tracking System Toward Automatic Recognition of Speech.
Signal Processing 2, No. 2, pp. 113-130, April, 1980.

[15] Lea, Wayne A., editor.
Trends in Speech Recognition.
Prentice-Hall, Inc, Englewood Cliffs, New Jersey, 1980.

[16] Lehiste, I.
Suprasegmentals.
The MIT Press, Cambridge, Massachusetts, 1970.

[17] Leung, H. C. and Zue, V. W.
A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech.
ICASSP, March, 1984.

[18] Lieberman, P.
Some Acoustic Correlates of Word Stress in American English.
The Journal of the Acoustical Society of America 32, No. 4, April, 1960.

[19] Mermelstein, P.
Automatic segmentation of speech into syllabic units.
Journal of the Acoustic Society of America 58(4), October, 1975.

[20] Morton, J. and Jassem, W.
Acoustic Correlates of Stress.
Language and Speech 8:159-181, 1965.

[21] Oshika, B. T., Zue, V. W., Weeks, R. V., Neu, H., and Aurbach, J.
The Role of Phonological Rules in Speech Understanding Research.
IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-23, No.1, February, 1975.

[22]
Seneff, S.
Pitch and Spectral Estimation of Speech Based on Auditory Synchrony Model.
ICASSP , March, 1984.

[23]
Shipman, D. W.
Development of Speech Research Software on the M.I.T. Lisp Machine.
103rd Meeting of the ASA , 1982.

Acoustic-Phonetic Constraints in Continuous Speech Recognition: A Case Study Using the Digit Vocabulary

by

Francine Robina Chen

B.S.E., University of Michigan
(1978)

S.M., Massachusetts Institute of Technology
(1980)

Submitted to the Department of
Electrical Engineering and Computer Science
in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology
June 1985

Signature of Author ..Francine R Chen..............................
Department of Electrical Engineering and Computer Science
May 22, 1985

Certified by ....
Victor W. Zue
Thesis Supervisor

Accepted by ....
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

Acoustic-Phonetic Constraints in Continuous Speech Recognition: A Case Study Using the Digit Vocabulary

by

Francine Robina Chen

Submitted to the Department of Electrical Engineering and Computer Science
on May 22, 1985 in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Abstract

Many types of acoustic-phonetic constraints can be applied in speech recognition. Shipman and Zue proposed an isolated word recognition model in which sequential constraints are applied at a broad phonetic level to hypothesize word candidates. Detailed acoustic constraints are then applied on a subsequent phone representation to determine the best word from the remaining word candidates. This thesis examines how their model can be extended to continuous speech. We used the recognition of continuously spoken digits as a case study.

We first conducted a feasibility study in which words and word boundaries were hypothesized from an ideal broad phonetic representation of a digit string. We found that strong sequential constraints exist in continuous digit strings and used these results to extend the Shipman and Zue isolated word recognition model to continuous speech.

The continuous speech model consists of three components: broad phonetic classifier, lexical component, and verifier. These components have been implemented for the digit vocabulary for the purpose of exploring how acoustic-phonetic constraints can be applied to natural speech. The broad phonetic classifier produces a string of broad phonetic labels from a set of parameters describing the speech signal. The lexical component uses knowledge about statistical characteristics of the output produced by the broad phonetic classifier to score each of the word hypothesis. Evaluation of this part of the system suggests that it can prune unlikely word candidates effectively.

Nine acoustic features were defined to characterize phones for verifying each of the word candidates. Evaluation of the verifier on the digit vocabulary demonstrates the power of a phone-based representation and of using a few well-motivated acoustic features for describing phones in an acoustic-phonetic approach. In addition to examining the application of speech constraints, evaluation of each of the components indicates that an acoustic-phonetic approach is potentially speaker-independent.

Thesis Supervisor: Victor W. Zue
Title: Associate Professor of Electrical Engineering and Computer Science

Acknowledgments

Many people have contributed to this thesis. I especially want to express my appreciation and thank:

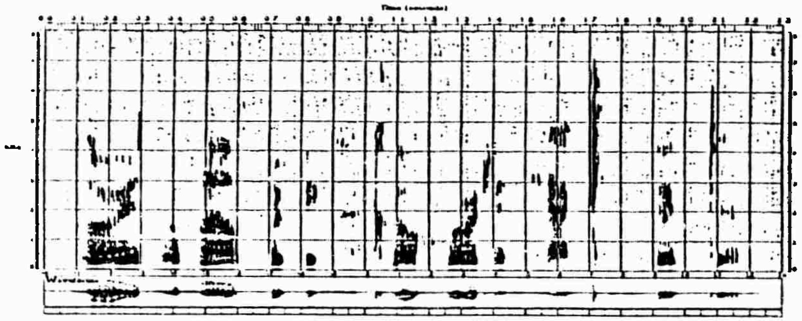
My thesis advisor, Victor Zue, for his guidance, support, friendship, and many Chinese meals. His insight and enthusiasm, and his ability to help me clarify my thoughts and encourage me to try new directions have made working with him a very rewarding experience;

My readers, Jonathon Allen and Kenneth Stevens, for their interest in this work and for their helpful comments on this thesis;

Kenneth Stevens for his advice while I've been a doctoral student and for making the Speech Communication Group a family;

Scott Cyphers, Jim Glass, Lori Lamel, Hong Leung, Mark Randolph, and Stephanie Seneff for discussions which helped to shape this thesis;

Scott Cyphers for helping me develop the rule interpreter, and for taking good care of the bears (our Lisp Machines);



Past and present members of the Speech Communication Group for support and for providing a pleasant and stimulating research environment;

My parents, Robin and June Chen, for their love, support, and understanding;

Leigh, for his love and patience.

Research support for this thesis has been provided by a Vinton-Hayes Fellowship and by contracts from DARPA, monitored through the Office of Naval Research, and by System Development Foundation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis examines an acoustic-phonetic approach to continuous speech recognition. The approach relies heavily upon low-level speech knowledge—knowledge about phonotactics, the lexicon, allophonic variation, and the duration of different speech units. In this thesis, the constraints provided by each type of low-level speech knowledge were studied and characterized. The constraint information was used to develop components of a speaker-independent, continuous digit recognition system as a research tool. Implementation of the system components allowed a better understanding of how speech constraints could be used in a recognition system.

Speech recognition by computers has possible applications in many areas, ranging from assembly line inspection to airline reservations to aids for the handicapped. Computer recognition of speech could simplify the interaction between humans and computers; one would only need to be able to talk in order to enter information into a computer. We would like speech recognition systems to be speaker-independent, so that a new user is not required to train a system before using it. Furthermore, we would like speech recognition systems to recognize continuous speech, as opposed to isolated words. We use continuous speech, not isolated words, when we speak; therefore, a continuous speech recognition system is more user-friendly. Continuous speech recognition systems have the added advantage that users could enter information into a computer more quickly, since the speaking rate is higher in continuous speech.

In the past, researchers have expended much effort developing and refining recognition systems based chiefly on engineering techniques. This is primarily a reflection of relatively primitive and incomplete knowledge about acoustic, phonetic, and other low-level characteristics of speech. However, we now understand these characteristics more fully than we did a decade ago. By exploiting what knowledge we have about speech and then pushing our knowledge further, better and more advanced speech recognition systems may be developed.

1.1 Speech and Speech Knowledge

Speech sounds are produced as air flows through and resonates in the vocal tract. Different speech sounds are due to different configurations of the vocal tract, each of which is associated with a set of resonant frequencies. In addition, different sounds are produced depending on the excitation source. The excitation may be at the glottis and/or at a constriction(s) in the vocal tract. When the excitation is at the glottis, the vocal folds can remain open to produce aspiration, as in /h/, or the vocal folds may vibrate rhythmically to produce voiced, periodic sounds, such as vowels and nasals. The peaks in the spectrum of a voiced sound are called formants and are labeled as F_i for the i^{th} formant. Thus, for example, the formant with the lowest frequency is called the "first formant" and is labeled F_1. When the excitation is at a constriction in the vocal tract, noisy aperiodic sounds (e.g., /s/, /š/) are produced. More than one source may be present during production of a sound. When both a noise and voicing source are present, voiced consonants (e.g., /v/, /z/) are produced Many speech scientists (e.g., Chomsky and Halle, 1968; Jakobson, Fant, and Halle, 1952) have described speech sounds in terms of these characteristics, that is, voiced or unvoiced characteristics, and other characteristics

such as frontness of a vowel.

Because the rate the vocal tract articulators can move is limited, the articulators are not instantaneously positioned to produce each sound. Consequently, each sound is affected by its neighbors, or context; this is called *coarticulation* (see for example, Heffner, 1950). As an example, there are at least two kinds of /k/; each is illustrated in the spectrograms of Figure 1.1. The particular realization of a /k/ depends on the adjacent vowel: The /k/ on the left is followed by a front vowel and is called a "front /k/," and the /k/ on the right is followed by a back vowel and is called a "back /k/." Note the higher burst frequency of the front /k/; this is due to the constriction formed with the hump of the tongue against the roof of the mouth being positioned farther forward in a front /k/ than in a back /k/.

Due to differences in rate of articulatory movement, some sounds are relatively stable in time compared to others. For example, an /n/ is much more stable than an /r/ (compare the /n/ and /r/ in Figure 1.2). The transition from vowel to nasal, primarily due to the velum lowering to couple the oral and nasal cavities, is rapid because the velum can be quickly lowered. Once the transition is made, the nasal is stable for its duration. In contrast, an /r/, produced by retroflexing the tongue, usually shows movement throughout its duration. Since the tongue cannot move as quickly as the velum, the time it takes to retroflex the tongue to produce an /r/ can be observed in a spectrogram as gradual lowering of F_3.

In many languages, only limited sequences of sounds are allowed (Sigurd, 1970; Shipman and Zue, 1982). For example, in an English syllable beginning with three consonants, the first consonant must be an /s/, the second either /p/, /t/, or /k/, and the third either /l/, /w/, or /r/. Furthermore, not all combinations of these sounds are allowed. Thus given a sequence of sounds, one can deduce whether or not it could be a word in a specified language.

The examples in this section have briefly introduced some low-level speech characteristics. These characteristics can be organized as low-level speech knowl-

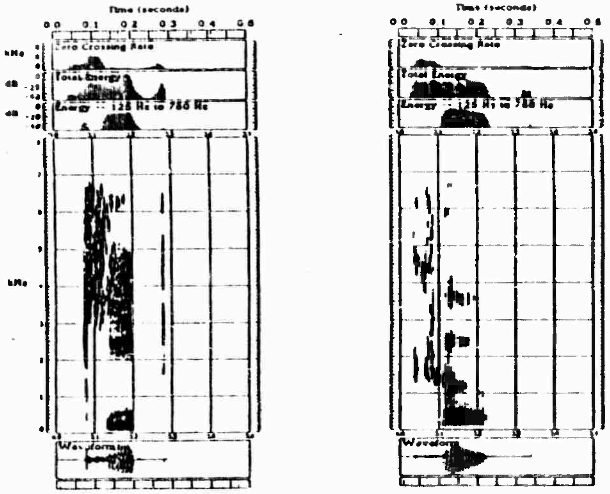


Figure 1.1: Front and back /k/'s: (a) front /k/ in "keep" (b) back /k/ in "coop"

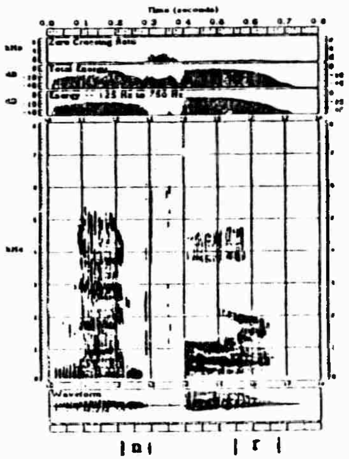


Figure 1.2: Rate of articulatory movement for /n/ and /r/ in the digit string "nine four"

edge: knowledge about acoustic characteristics of sounds, coarticulation, duration of sounds, and phonotactics. The formulation of this knowledge into speech constraints and the use of these constraints in recognition of natural speech were explored in this thesis.

1.2 Speech Recognition Systems

The speech recognition systems developed during the past 15 years have used varying amounts and types of speech knowledge. Some systems, such as those based on template-matching, use little, if any, speech knowledge. In contrast, the explicit use of speech knowledge forms the basis of the FEATURE system developed at Carnagie Mellon University. In this section, some benchmark recognition systems are described. (Many excellent reviews of major recognition systems have been written, such as by Lea, 1980 and Klatt, 1977.) Systems which use little explicit low-level speech knowledge are described first, followed by descriptions of systems which progressively use more speech knowledge in an explicit manner. At the end of this section, the use of speech knowledge by each of the described systems is compared.

1.2.1 Template Matching

Most speech recognition systems presently on the market are based on a mathematical approach which combines template matching with a time-alignment procedure known as dynamic time warping. These systems perform with over 95% accuracy on speaker-trained isolated-word and limited-vocabulary connected-word tasks (Kaplan, 1980; Doddington and Schalk, 1981). In template matching, each recognition unit, for example, a word, is represented by at least one template, created from a set of training utterances (Rabiner, 1978). Each template is composed of a sequence of patterns in time and each pattern, in turn, is some parametric representation of speech. For example, a template could be composed of the sequence of linear prediction coefficients sampled every 5 msec throughout a training utterance.

Dynamic time warping is a method based on dynamic programming for nonlinearly aligning two sequences. Some systems constrain the amount of time compression and expansion allowed; for example, Myers and Rabiner (1981a) limited time normalization to a ratio of 2:1 between the reference and test templates.

To recognize a word, an isolated word recognizer compares the input signal against each of the stored word templates using dynamic time warping and a predefined distance metric (e.g., Itakura's distance metric, see Itakura, 1975). The best alignment between the input signal and each template is found. The input word is classified as the word corresponding to the best matching template, that is, the word with the smallest distance.

Many constraints on isolated words are not present in continuous speech, and, as a result, extension of isolated word techniques to continuous speech is not straightforward. In isolated word recognition, the word endpoints are "known." In contrast, in continuous speech recognition, the word endpoints are unknown and coarticulation occurs between words. Because handling unknown endpoints requires extra computation, systems developed to extend template matching to continuous speech (e.g., Sakoe, 1979; Kato, 1980; Myers and Rabiner, 1981a; Myers and Rabiner, 1981b) have used very small vocabularies. Defining templates which differentiate among similar words is not an easy task using conventional template representations, such as linear prediction coefficients, because such representations do not adequately emphasize fine differences. Defining templates for continuous speech is an even more difficult task because one must have a framework to describe coarticulation across word boundaries. Conventional template representations do not lend themselves to descriptions of coarticulation because they represent an acoustic event and not an acoustic manifestation of a phonetic event. For reasons such as

these, extending template matching from isolated word recognition to less restricted speech recognition tasks has proven to be difficult.

1.2.2 Harpy

The HARPY system (Lowerre, 1977; Lowerre, 1980) demonstrated the best performance of all the continuous speech recognition systems developed under the ARPA project[1], with less than 5% semantic error on sentences from a highly constrained grammar. Quasi-stationary segments derived from simple parameters, called "zapdash" parameters, were "mapped to the network states based on the probability of match...by use of phonetic templates" (Lowerre 1977). Along with juncture rules, 98 templates were used to represent all possible allophones and speaker variations. A score was assigned to each node in the network based upon how well the sapdash parameter values in a segment of speech matched the templates. Finally, the finite state graph was searched using a heuristic but efficient search algorithm, called a beam search, to find the best path (subject to constraints used in the algorithm) through the network and the corresponding best sentence.

The developmental effort in Harpy concentrated on using high-level knowledge. As a result, the Harpy system relies heavily on higher level language constraints, such as its constrained grammar. Innovations introduced in the system include higher level processing of the input signal, a precompiled network embodying many forms of higher level knowledge, and the beam search algorithm. In contrast to its well-developed high level processing, Harpy's relatively primitive front end performed only rough segmentation based upon zero crossing rates and peaks in smoothed and differenced waveform parameters (the "zapdash" parameters).

The success of the Harpy system demonstrates that higher level constraints are

[1] During the early seventies, the Advanced Research Project Agency (ARPA) sponsored a five year project, involving approximately ten research groups, to study the feasibility of building systems for speaker-independent, continuous speech understanding

useful in speech recognition. In addition, it illustrated how constraining a task can be used to reduce a problem to manageable proportions. However, using only higher level constraints can reduce the task too much to be of general use. In fact, a general criticism of Harpy is that its grammar is so constrained that it is not habitable. Even so, the Harpy system exemplified how higher level language constraints can be applied to the recognition problem.

1.2.3 IBM

IBM has developed two benchmark systems: a speaker-trained continuous speech recognition system and a speaker-trained isolated word recognition system. The IBM continuous speech recognition system (Jelinek, 1975; Jelinek, 1976; Jelinek, 1981) has a recognition accuracy of 91% correct (Jelinek, 1980) on words contained in sentences in the 1000 word vocabulary of the *Laser Patent Text*. The isolated word recognition system has a recognition accuracy of 95% on a 5000 word office correspondence vocabulary (Bahl et al., 1983).

Both systems are based upon Hidden Markov Modeling, a statistical technique which IBM has applied to model specified speech units. In both systems, each word in the lexicon is represented as a sequence of phonemes in a finite state graph. The phonemes, in turn, are represented as a sequence of templates which attempt to capture variations of all phonemes (allophones) in the language. In the continuous speech recognition system, the templates are computed from DFT (Discrete Fourier Transform) coefficients, which evenly weight the spectrum even though the information content in the spectrum is not uniform. In contrast, in the isolated word recognition system, principal component values derived from the DFT representation are used as input from the front end, weighting the DFT coefficients in accordance with characteristics of speech sounds.

Hidden Markov Modeling characterizes the speech signal by statistical methods; thus, the system can be trained without human input. However, Hidden Markov

Model systems require large amounts of speech data and computation for training. In particular, training a continuous speech recognition system to one speaker may require many hours of speech and many hours of computer time, making it unacceptable for general use. In contrast, the isolated word recognition system requires much less training than the continuous recognition system, although it still does require hours of computer time on IBM's largest computer. By constraining the task to isolated word recognition, IBM made the training and computational costs for a Hidden Markov Model recognition system manageable.

1.2.4 Hearsay II

The research effort of CMU's Hearsay II system (Erman and Lesser, 1980; Lesser et al., 1975), developed under the ARPA project to recognize continuous speech from several speakers, was directed at developing and studying the interactions of knowledge sources Each knowledge source contained "knowledge," such as speech descriptions, needed by the recognition system to solve a particular recognition task. The knowledge sources were modular, which allowed easy modification of knowledge, but the cost was slower recognition and a more complex control structure. In the resulting system configuration, independent, parallel, knowledge sources communicated through a multilayer global blackboard The "layers" of the blackboard included: segment labels, syllables, proposed lexical items, accepted words, and partial phrase theories. A knowledge source was activated when new information on the blackboard caused its preconditions to be met. An activated knowledge source would attempt to provide information to a higher level (bottom-up analysis) or lower level (top-down analysis).

Much less emphasis was placed on development of low-level knowledge sources concerned with acoustic-phonetics relative to higher level knowledge sources concerned with syntax and semantics. For example, the Hearsay II segmentation component simply used template matching and a well-developed algorithm (Itakura's distance metric) to assign to each segment a label corresponding to one of 98 possible templates. The performance of the low-level knowledge sources was poor: the segmentation component assigned the correct label as its first choice only 42% of the time, and the word hypothesizer hypothesized the correct word to be within the top 50 candidates out of a possible 1000 words only 70% of the time (Klatt, 1977). But higher level syntactic and semantic knowledge sources allowed recovery from these errors by top-down word hypothesization. Thus, Hearsay II demonstrated the utility of knowledge to drive a recognition system, especially the use of higher level knowledge sources.

1.2.5 HWIM

The HWIM system (Hear What I Mean), developed under the ARPA project at Bolt, Beranek, and Newman, had fixed components in a roughly hierarchical structure. The system initially segmented and labeled the speech signal as a set of phonetic transcription alternatives arranged in a "segment lattice." The purpose of the lattice was to avoid fatal segmentation errors by allowing ambiguity. The first choice accuracy of assigning the correct phonetic label (out of 71 possible labels) was only 56% (Schwartz and Zue, 1976).

The designers of HWIM introduced several interesting ideas for speech recognition. These included word verification at the parametric level, phonological rules for building a lexical decoding network, and top-down verification using the context and position of a word (Cook and Schwartz, 1977). HWIM's parametric word verifier scored word candidates at the parametric level by matching frames of a word candidate with frames of the corresponding synthesized word. Speech knowledge was used in developing the front end descriptors for each of the phonetic labels and in developing phonological rules. Language knowledge was used to find the best path through the word lattice and in top-down verification. However, the methods used to incorporate speech knowledge into the system were not fully explored. For

example, "intuitive human guesses, not statistically measured estimates" of likelihoods for word pronunciations were used to evaluate the system due to lack of time (Wolf and Woods, 1980). This was done because obtaining sufficient statistics to produce meaningful pronunciation likelihoods is difficult.

1.2.6 FEATURE

The FEATURE system (Cole et al., 1982), which recognized the letters of the alphabet, used acoustic features of speech for discriminating among speech sounds. To recognize a word, the system extracted acoustic features from parameters between four temporal anchor points. These anchor points were chosen to take advantage of the monosyllabic structure of most of the lexical items. The parameters used in the system, motivated by acoustic-phonetic knowledge of speech, include formant frequencies, fundamental frequency of voicing, sero crossing rate, total energy, low frequency energy, mid-frequency energy, and high frequency energy.

FEATURE's average recognition rate was 89.5% correct when tested on 10 male and 10 female speakers. In light of the difficult vocabulary (many of the letters of the alphabet sound similar), FEATURE's performance demonstrates that acoustic-phonetic knowledge can be useful in speech recognition. However, extensibility of the FEATURE algorithm to continuous speech is uncertain. The recognition algorithm took advantage of the monosyllabic nature of the vocabulary words and the analysis was done from four anchor points. In addition, many hours of a speech scientist's time were required to develop features for identification in the limited vocabulary; many more hours would be needed to develop features for all contexts.

1.2.7 Speech Knowledge in Recognition Systems

We have seen several approaches to speech recognition, each using varying amounts of speech knowledge. Template matching techniques use constraints to define a manageable task (e.g., speaker-trained and isolated word tasks) and are relatively easy to develop because they are mathematically well defined. In template matching systems, only a very small amount of speech knowledge is used relative to the potential amount of knowledge that could be used. In these systems, speech knowledge is usually incorporated by limiting the amount of time compression/expansion, reflecting some knowledge about limits on speech rate variation. Speech knowledge has also been incorporated through choice of recognition unit; for example, Rosenhurg et al. (1983) developed a system which uses the demisyllable as the recognition unit.

The Harpy system used more speech knowledge than template matching systems. Some speech knowledge was needed to develop the "zapdash" parameters, and template selection required some speech knowledge. However, most of the knowledge used was higher level language knowledge, such as grammar and syntax.

The IBM systems showed how some speech knowledge combined with well defined mathematical techniques can successfully be used for recognition. Speech knowledge was explicitly used in defining the sequence of templates representing a word and in defining allowable word pronunciations due to coarticulation. Statistical characterization of the speech signal provided much strength to the systems. However, this characterisation was performed without explicit use of knowledge about speech and also required a lot of training data.

The importance of language constraints in speech recognition and how such constraints can be used to recover from low-level errors was demonstrated by Hearsay II. HWIM and FEATURE illustrated that low-level speech constraints may be important in speech recognition. In particular, HWIM exemplified how constraints on phonological variations may be important, and FEATURE exemplified how acoustic-phonetic constraints may be important.

Each of these systems has contributed to our understanding of how to use speech knowledge in speech recognition. However, we still need to understand better the constraints provided by different types of speech knowledge, especially low-level

speech knowledge, and how to cohesively use the constraints in the speech signal for recognition.

1.3 Problem Statement and Overview

This thesis studies the application of low-level constraints in the speech signal to continuous speech recognition, particularly the task of recognizing continuous digits. Lexical, durational, acoustic, phonetic and allophonic speech constraints are examined and the utility of these constraints is tested on natural speech. In contrast, high level knowledge about the language, such as grammar, syntax, and semantics, is not addressed. The investigation was divided into three parts:

1. develop a continuous speech recognition model based upon constraints in the speech signal

2. implement components of the model, making the required modifications to accommodate variabilities in the speech signal

3. explore the use of detailed acoustic analysis of phones for verification of word hypotheses

Chapter 2 develops a continuous speech recognition model which relies heavily upon speech kuowledge and is based on seqeuential constraints, as used by Shipman and Zue (1982). Shipman and Zue's work in sequential constraints for isolated word recognition is described first. A feasibility study which was conducted to show that strong sequential constraints exist at the broad phonetic level in digit strings is described next. The results of the study indicate that for a continuous speech recognition task such as the digits, a recognition system can initially process continuous speech at the more robust broad class level, rather than at the detailed level of the benchmark systems. This ph losophy was used in the development of the recognition model.

Other low-level speech constraints which may be useful in a recognition system are described next. The use of additional constraints provided by the low-level properties of the speech signal performs two functions. First, the additional constraints "counteract" the loss of word endpoint constraint in continuous speech. Second, the additional constraint may allow "higher" level language constraints to be relaxed. For example, the grammatical constraints in Harpy may be relaxed to form a more habitable grammar.

In the last section of Chapter 2, the Shipman and Zue model is extended to continuous speech. The general organization of the model is presented, and the incorporation of constraints in the speech signal into the model is discussed.

The next two chapters describe the implementation of the components in the recognition model. Chapter 3 considers how speech constraints can be applied in the broad phonetic classifier and lexical component. Chapter 4 explores an acoustic-phonetic approach to verification of word hypotheses in a word lattice.

In Chapter 3, refinements to the continuous speech recognition model to handle the variations which occur in natural speech due to interspeaker and intraspeaker variations are described. These variations can lead to recognition errors unless the recognition algorithm is developed to explicitly deal with them. In addition, contextual and coarticulatory variations occur in natural speech and must be incorporated into the algorithm. For example, a person may pronounce "five" with or without a /v/, as in [fa^{y}ve^{y}t$^\urcorner$] ("five eight") and [fa^{y}na^{y}n] ("five nine"), respectively A method for segmentation and labeling by characterizing speech in terms of acoustic features by the broad phonetic classifier is described. Then an algorithm for applying sequential constraints to natural speech using knowledge of front end characteristics is developed. Finally, an investigation of the application of path, allophonic, and durational constraints is presented.

Chapter 4 describes how the knowledge gained from low-level constraints is used when performing detailed acoustic analysis. In particular, verification of word hypotheses using a phone-based representation was explored. A set of acoustic features for characterizing and identifying the phones in the digit vocabulary are described. A method for using information from the features to score each phone hypothesis is presented and then the method is evaluated.

Each of the implemented components was evaluated. The broad phonetic classifier was evaluated by comparing its output to a hand-labeled phonetic transcription. The lexical component was evaluated using output from the broad phonetic classifier, since the main contribution of the lexical component is in using the characteristics of the broad phonetic classifier in evaluating word hypotheses. The verification component was evaluated through incremental simulation, so that errors due to the verifier could be separated from errors due to other components.

Chapter 5 justifies the assumptions made in the thesis and outlines the contributions of the thesis. This chapter discusses the most prominent assumption, that an acoustic-phonetic approach has many advantages over other approaches. Additionally, the characteristics of the digit vocabulary and how the choice of vocabulary affects the study are described. The utility of a preprocessor, especially an acoustic-phonetic preprocessor, the use of segments and sequential constraints in recognition, and computational considerations are also discussed. Finally, the contributions of the thesis are outlined and suggestions are made regarding ways the research can be refined and extended.

Chapter 2

A Speech-Knowledge Based Recognition Model

This chapter discusses a continuous speech recognition model which uses speech knowledge to constrain the recognition task during each processing step. The philosophy of the model is general enough to serve as a basis for developing large vocabulary continuous speech recognition systems using low-level speech knowledge. Since the model was implemented on a limited task, the digits, some of the background work is based on analysis of only the limited vocabulary. Suggestions on how the model may be relevant to larger vocabularies are discussed in Chapter 5.

The model follows the work of Shipman and Zue, described in Section 2.1, on sequential constraints in isolated words. The feasibility of a sequential constraint based approach for continuous speech was analyzed with a modeling experiment, described in Section 2.2. This study showed that strong sequential constraints on a broad class representation of continuous speech can be used to specify word hypotheses and corresponding word boundaries for the limited case of digit strings. However, sequential constraints in continuous speech do not provide as much constraint in identification of the utterance as sequential constraints in isolated words, because definite word endpoints are unknown; consequently, other constraints were

used in recognition. In Section 2.3, different types of speech knowledge and how each can be used as a recognition constraint are discussed. In Section 2.4, a recognition model which is based upon conclusions from the feasibility study and which uses constraint information is described. The general structure of the model is described first, followed by a discussion of how speech constraints should be used in each component of the model.

The model performs initial analysis at the broad phonetic level. Sequential constraints are applied to produce word candidates and then each word candidate is scored using more detailed phonetic analysis. The best sentence is recognized as the best scoring sequence of word candidates. The proposed model for continuous speech recognition incorporates the speech constraints described in Section 2.3 (in addition to sequential constraints) to further specify the use of speech knowledge in the Shipman and Zue model and to extend the model to continuous speech.

2.1 Sequential Constraints in Isolated Words

House and Neuburg (1977) introduced the idea of representing an utterance as a sequence of broad phonetic classes. They introduced this idea with the belief that gross linguistic categories could be identified more easily than a detailed phonetic representation. In 1982, Shipman and Zue conducted a study on sound patterns in isolated words. The results of the study demonstrated that the sound patterns of English impose strong sequential constraints on the words in the language. For example, they found that the only word in Webster's 20,000-word pocket dictionary satisfying the template:

|consonant| |consonant| |l| |vowel| |nasal| |stop|

is "splint," illustrating that sequential constraints exist even when some phonemes are represented as a broad phonetic class (relaxing the constraints). Shipman and Zue performed an experiment in which each sound was represented as one of six broad phonetic classes: strong fricative, weak fricative, vowel or syllabic consonant, stop, nasal, and liquid or glide. The average number of words matching a particular sequence, normalized by frequency of occurrence in the Brown Corpus, was reduced to approximately 0.2% of the lexicon when this representation was used, and the maximum cohort size was about 1% of the lexicon.

These results show that strong sequential constraints exist in broad phonetic representations of words. Furthermore, since more detailed distinctions must be made to produce a detailed phonetic representation than a broad phonetic representation, a broad phonetic representation is more robust than a detailed phonetic representation. Therefore, sequential constraints at the broad phonetic level also should be more robust. Thus, Shipman and Zue proposed the following approach to isolated word recognition. The sound units are first classified into several broad categories which can be determined with little error. Then, indexing into the lexicon, a subset of the lexicon is found as word candidates. Finally, a detailed analysis of acoustic differences is used to recognize the word.

2.2 Sequential Constraints in Continuous Speech

The Shipman and Zue study demonstrated that strong sequential constraints exist on words in the English language. If strong sequential constraints were shown to exist in continuous speech, then the recognition approach outlined by Shipman and Zue could be extended to continuous speech. To test this hypothesis, a feasibility study of sequential constraints in continuous speech was conducted on a limited vocabulary, the digits from "zero" through "nine," with the idea that the task may later be expanded if the initial approach proved viable. In this section, the results of this study are outlined.

In the isolated digit vocabulary, there are two unique consonant clusters, /θr/ and /ks/, but in connected speech the word endpoints are not obvious. Thus,

consonant "clusters" can be formed from combinations of word-final consonants with word-initial consonants. In the digit vocabulary, there are 32 (ignoring gemination) unique sequences of a word-final consonant followed by a word-initial consonant. However, none of these consonant sequences are allowable within a digit. Since the set of consonant sequences at word boundaries and within a digit are mutually exclusive, all word boundary locations between two consonants in an ideal phonetic transcription can be determined by examining phoneme sequence pairs and using constraints on allowable consonant sequences.

In a broad phonetic representation, there is only one non-unique phoneme class sequence in the digit vocabulary: [stop] [strong-fricative], as in "six" and "eight seven." Thus, sequential constraints in digit strings should still exist at the broad phonetic level. This result is important because a broad phonetic representation is more easily and robustly derived automatically from the speech signal than a phonetic transcription where inter-speaker variations may be of the same magnitude as phonetic differences. In addition, such a representation may be more robust against environmental variability.

The application of detailed phonetic and broad phonetic sequential constraints were examined more qualitatively. Sequential constraints in the digit vocabulary were found to be strong enough to uniquely parse a detailed phonetic transcription of a digit string to recover the original digits. In contrast, broad phonetic sequential constraints are not as strong. An experiment was conducted to examine the application of broad sequential constraints to an ideal broad phonetic representation of digit strings for identifying individual digit boundaries from a string. The representation was ideal because word boundary effects were ignored and the broad phonetic transcription was correct. The constraints were expressed as sequences of broad phonetic classes which could represent a digit. In this study sequential constraints were used to propose words *and* corresponding word boundaries in digit strings; in contrast, in the Shipman and Zue study, sequential constraints were used

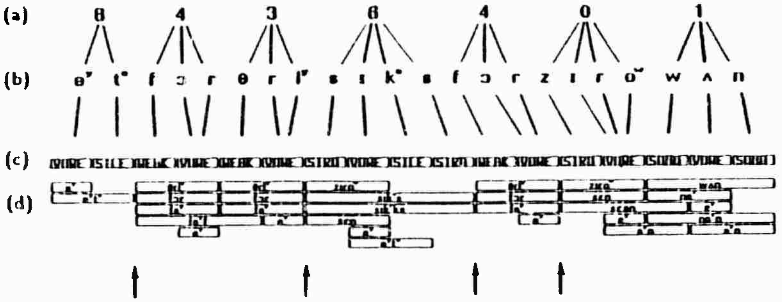


Figure 2.1. Example of Application of Sequential Constraints to an Ideal Broad Phonetic Representation

to reduce the number of cohorts associated with a word.

Ten thousand digit strings containing 34,947 word boundaries were used in the study. The digit strings were of random length and were composed of digits in random order. A phonetic representation of each string was produced by concatenating a phonetic transcription of each digit in the string. The phonetic transcription of each digit was randomly selected each time from the set of transcriptions observed for the digit in a transcribed set of training utterances, thus allowing for multiple pronunciations of a digit. The phones in the phonetic transcription were then mapped into six broad phonetic categories (strong fricative, weak fricative, silence, vowel, souorant[1], or short voiced obstruent) to produce a broad phonetic transcription.

An example of the mapping procedure is shown in Figure 2.1. The digit string "8436401" Figure 2.1a is mapped into a phonetic string (b), and then mapped into a broad phonetic string (c). All the words with a broad phonetic representation which matches a portion of the segmentation string are shown in (d). The word boundaries marked by the vertical arrows can be identified with certainty because

[1] In this thesis *sonorant* refers to a consonant class, rather than a distinctive feature.

all word hypotheses in the region either begin or end at the segment boundary. Two word boundaries cannot definitely be identified. For example, the boundary between the first "four" and the "three" cannot be definitely identified because the word [fa^{y}f] spans the boundary. Using this representation, 66% of the word boundaries were found. Thus strong sequential constraints exist in digit strings even at the ideal broad phonetic level.

The ability to identify 66% of the word boundaries in ideal data implies that sequential constraints can be useful in hypothesizing word candidates from a broad phonetic representation. However, these results are not directly applicable to real data. In real systems, the lexical access component must tolerate phonetic variability in the signal and front end errors. This reduces the strength of the sequential constraints and may increase the number of word candidates. To help reduce the number of word candidates, other speech constraints can be used.

2.3 Speech Knowledge for Speech Recognition

Speech can be characterized in many ways—acoustically, phonetically, by sequential ordering of sounds, and by duration of sounds. Knowledge about these speech characteristics can be used as constraints in recognition. Sequential constraints examined by Shipman and Zue were expressed using several representations, including detailed phonetic and broad phonetic representations. Similarly, other types of speech constraints can be expressed at both the phonetic and broad phonetic levels. These constraints can be used in many areas of recognition, such as segmentation and labeling the speech signal to produce a sequence of broad class labels. Different types of speech information which can be used as constraints in speech recognition at the phonetic and broad phonetic level of description are described in the following sections.

Appropriate points in the recognition process for applying each constraint are also discussed. A constraint should be applied when it is most effective and when the system has enough knowledge to check if the criteria for the constraint to be applied are present. Before applying a constraint, the amount by which a particular constraint reduces a task should be considered. For example, when both phonetic and broad phonetic constraints provide sufficient constraint to be used, broad phonetic constraints are applied first. Constraints defined at the broad phonetic level require less detailed knowledge and are dependent upon more robust information. Furthermore, the additional information gleaned by using these constraints may allow better use of detailed phonetic knowledge.

2.3.1 Acoustic and Phonetic Knowledge

Speech scientists have developed a set of distinctive features for characterizing speech sounds (e.g., Jakobson et al., 1952). Some of these features (such as whether a vowel is high or low) have well defined acoustic correlates. Other features (such as whether a consonant is distributed) do not have obvious acoustic correlates (Fant, 1969) and therefore do not lend themselves well for use in recognition systems. Speech can also be described by a set of acoustic characteristics. Spectrograms[2] are one representation in which acoustic-phonetic information can be observed. In Figure 2.2, vowel regions are indicated by the bars underneath the spectrogram. Note the striations in the signal and presence of energy below 800 Hz in these regions. Properties of speech sounds which can be described acoustically will be defined to be *acoustic features* in this thesis. For example, a "large amount" of energy below 800 Hz satisfies this definition of an acoustic feature.

Properties of large sound classes can be described using acoustic features. For example, one property of vowels and voiced sonorants is that they are strongly voiced, and this can be described by acoustic features such as a "large amount" of

[2] A spectrogram is a frequency versus time representation of a signal where amplitude is represented by print darkness.

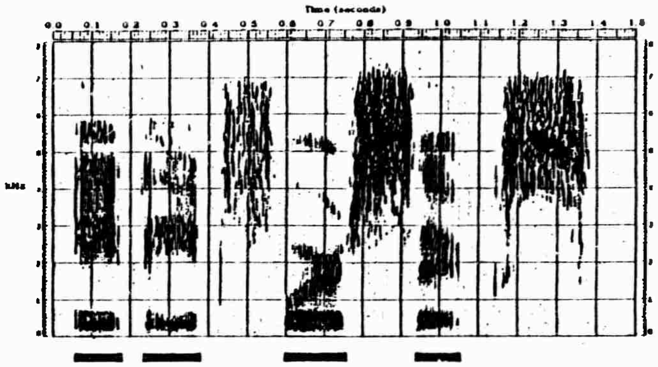


Figure 2.2: Sample Spectrogram. Bars indicate vowel regions

energy below 800 Hz. Acoustic features can also describe more detailed acoustic events, such as a rising second formant or a sharp onset. These detailed acoustic features are useful in making fine phonetic distinctions, as between a /θ/ and the release in /t/.

An acoustic feature is defined using acoustic constraints to describe a characteristic of speech. The combination of acoustic features describing a class of speech sounds define a *phonetic constraint*. For example, a strong fricative is characterized by a non-periodic energy source and a large amount of high frequency energy (an acoustic feature). The aperiodic signal, in turn, is characterized by a high zero crossing rate. Very low frequency energy may also be present in voiced strong fricatives, especially during the initial portion. The acoustic features of a large amount of high frequency energy, a high zero crossing rate, and maybe low frequency energy at the beginning of a region form one description of the strong fricative class of speech sounds and define a phonetic constraint for the strong fricative class.

Since phonetic constraints are defined by acoustic features, acoustic constraints are applied before phonetic constraints. Coarse acoustic and broad phonetic constraints, defining broad classes should be applied early in the recognition process because these robust descriptors provide strong constraints in narrowing down the word candidates. In addition, these constraints require little computation and no contextual knowledge. Detailed acoustic and phonetic constraints defining phones should be applied later in the recognition process. The realization of a phone is context dependent, thus how "good" a phone candidate is can be more accurately accessed given the context. Since the context is not available until at least one pass is made over the portion of the signal being considered, a recognition system does not have enough knowledge to apply detailed acoustic and phonetic constraints early in the recognition process.

2.3.2 Lexical Knowledge

Lexical information is another type of speech knowledge which can be used to constrain the recognition problem. In the English language, many patterns exist in any subset of words. As the Shipman and Zue results showed, only a small percentage of words can be represented as a particular sequence of broad phonetic classes. Thus the limit on the number of words associated with a particular sequence is one expression of lexical constraint in isolated words.

Lexical constraints are expressed in continuous speech a couple of ways. Each word in the lexicon can be represented as a sequence of broad classes or phonemes. Based on these sequential constraints, word boundaries can occur only in positions where a sequence matches a portion of the segmentation string. Once words are proposed, the correct word sequence should form a path through the lattice of proposed words. Path constraints, which are another expression of lexical constraints, can be used to prune words when at least one of the following conditions is not satisfied: 1) a preceding adjacent word exists or the word is sentence initial; 2) a following adjacent word exists or the word is sentence final. Paths which traverse these word candidates form an incomplete path in the word lattice. Pruning these

candidates reduces the computation needed in further processing.

In a recognition system, sequential constraints can be applied at the broad phonetic or detailed phonetic level to propose word candidates. Shipman and Zue's work and the results of the feasibility study discussed in Section 2.2 indicate that much constraint is available at the broad phonetic level. Since an accurate broad phonetic segmentation is more easily computed than an accurate detailed phonetic segmentation, word candidates should initially be proposed from a broad phonetic representation. Application of path constraints follows naturally, using knowledge of the endpoint locations of proposed words in the word lattice to prune word candidates which lead to a "dead end" path.

2.3.3 Knowledge of Duration

Durational constraints may be expressed in segment, word, and other representations. These constraints are derived from knowledge that the duration of a given speech unit is limited to a particular range. Durational constraints can be used to rule out a hypothesized unit which has a duration outside the observed range of the unit. For instance, a segment may represent one or more phones. In the digit string "one seven," a strong fricative segment is used to represent one phone, the /s/ in "seven." But in the phrase, "six seven," the /s/ in "seven" can share the same strong fricative segment as the final /s/ in "six"; this strong fricative segment represents two phones. In this example, durational constraints may be used to rule out the possibility that only one phone is represented by the strong fricative segment or the possibility that two phones are represented by the strong fricative segment. Figure 2.3 shows the distribution of the duration of a model segment representing one phone versus the distribution of the duration of the same model segment when it represents two phones. In region (a), only one phone is possible so a model segment with a duration in region (a) would not be allowed to represent two phones; similarly, in region (c), only two phones are possible and a model segment with a

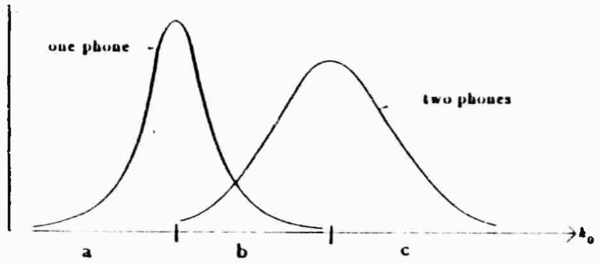


Figure 2.3: Schematized Histogram of Segment Duration for: (a) one phone (b) one or two phones (c) two phones

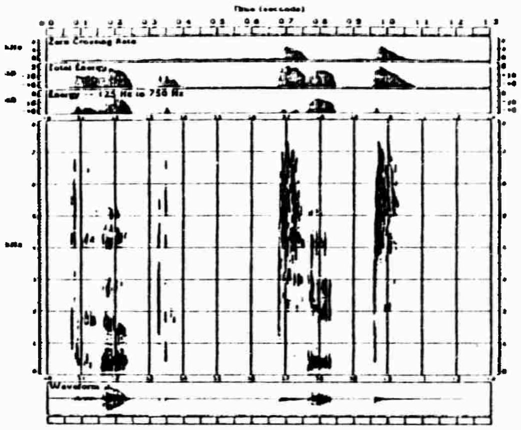
duration in region (c) would not be allowed to represent one phone; and in region (b), one or two phones are possible so neither can be ruled out.

Durations of words may also be used as a constraint. However, durations must be used with care. Speaking rate varies from sentence to sentence and also within a sentence. Consequently, unreasonable word durations must be determined using durational information from at most the sentence being recognized.

Preconditions for applying durational constraints depend only upon the presence of the unit being considered. Since durational constraints are used to *rule out* units for which the duration is unreasonable, these constraints should be applied soon after the unit is hypothesized and before further processing. For example, durational constraints on the number of phones represented by a segment can rule out some phone combinations with little computation. Detailed acoustic analysis may also be used to determine whether a segment better represents a single phone candidate or a pair of phone candidates. Since this is more computationally expensive, reducing the number of candidates by early application of durational constraints is preferred.

2.3.4 Knowledge of Allophonic Variation

Depending upon the context of a phoneme, many different realizations of that



(a) (b)

Figure 2.4: Spectrogram of /u/ in: (a) the word "poop" (labial context) and (b) the word "toot" (alveolar context)

phoneme are possible. For example, the second formant in /u/ is much lower in frequency when surrounded by labial consonants than when surrounded by alveolar consonants, as illustrated in the spectrograms of Figure 2.4. Allophonic constraints are based upon this type of knowledge. Allophonic constraints would specify that a hypothesized /u/ with a low F_2 in the context of alveolar consonants be ruled out as a candidate phoneme, but that a hypothesized /u/ with a high F_2 in the context of alveolar consonants is alright. Thus, if an /u/ is hypothesized, but F_2 is high and labial consonants are known to surround the vowel, then /u/ can be ruled out as a candidate phoneme.

Allophonic constraints can also be specified at the broad phonetic level. Different allophones of /t/ are used in pronunciations of the word "eight." Three different pronunciations of "eight" and the context, if required, for each pronunciation are shown in the first two columns of Table 2.1. Note that "eight" may be pronounced with a released or unreleased /t/ in any environment; but "eight" is pronounced with a flapped /t/ only when the following word begins with a vowel. Thus an

Table 2.1: Pronunciations of /t/

Transcription	Context	Broad Class Representation
[eʸt˺t]		vowel silence fricative[3]
[eʸt˺]		vowel silence
[eʸɾ]	__vowel	vowel short-voiced-obstruent

"eight" pronounced with a released or unreleased /t/ has no contextual constraint, but an "eight" with a flapped /t/ can only be followed by a word which begins with a vowel; if none of the following words begins with a vowel, then the "eight" should be removed as a word candidate. The rightmost column of Table 2.1 shows how the three different pronunciations can also be represented using broad phonetic classes. We see that the allophonic variations which occur in /t/ can be expressed at the broad phonetic level; broad phonetic allophonic variation can also be expressed for some other consonants. Hence, allophonic knowledge at the broad phonetic level can be used to rule out word candidates when some consonant contexts are incompatible. Allophonic knowledge at the phonetic level can be used to rule out word candidates based upon more subtle differences, such as the vowel realisations illustrated by the earlier /u/ example.

As previously stated, broad phonetic constraints should be used before phonetic constraints when both types of constraints are used. Thus, allophonic knowledge at the broad phonetic level is used first to prune the word lattice when a broad phonetic representation of each word candidate is available. Allophonic constraints at a detailed phonetic level are used later to discriminate between word candidates based upon fine differences in similar speech sounds.

[3] The release of a /t/ actually consists of a burst followed by aspiration. In this thesis, the broad class of "fricative" was generalized to include aspiration in addition to fricative sounds.

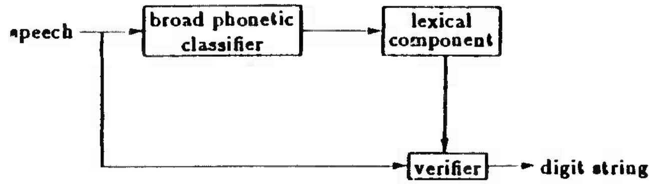


Figure 2.5: Phonetically-based continuous digit recognition system

2.4 Recognition Model

In this section, an acoustic-phonetic continuous speech recognition model is developed. The model's particular configuration was motivated by the results of the feasibility study described in Section 2.2 and the philosophy of using speech knowledge to constrain the recognition task.

In Shipman and Zue's proposed model for isolated word recognition, the speech signal is represented as a sequence of broad classes from which word candidates are hypothesized. Each word candidate is then analyzed in more detail and the word is verified or rejected. This same idea can be used for continuous speech. From a broad class representation of a sentence, word candidates and also the corresponding endpoints are hypothesized, producing a lattice of words. Each word candidate in the lattice is analyzed in more detail and then scored. Finally, the best sequence of words spanning the lattice is found.

The main components of this system, shown in Figure 2.5, are a broad phonetic classifier, lexical component, and verifier. The broad phonetic classifier produces a broad class segmentation of the incoming signal based on coarse characteristics of the parameterized signal. The use of coarse characteristics by this component makes its output potentially robust with regard to speaker variability. The broad class segmentation is then input into the lexical component where the phonetic transcription of each word in the lexicon is matched against the segmentation produced by the system. This yields a lattice of word candidates; the lattice is further reduced using lexical, allophonic, durational, and contextual constraints. When more than one word candidate exists over a portion of the reduced lattice, the verifier will rank order the candidates based on detailed acoustic analysis and knowledge about feature characteristics of the phones in each word candidate.

Speech constraints are applied at appropriate points in processing of the system. The most general constraints are applied first, with each new constraint being more specific. The constraints used by the broad phonetic classifier to produce a broad phonetic representation are: coarse acoustic features of speech sounds, broad phonetic characteristics, and durational constraints on broad phonetic classes. The broad phonetic classifier first derives coarse acoustic features from the speech signal. Broad phonetic classes are hypothesized based upon the coarse acoustic features present and broad phonetic constraints. Once broad phonetic classes are hypothesized, durational constraints are applied to check that the duration of a segment is reasonable. For example, a short voiced obstruent, such as an intervocalic /v/, should be of shorter duration than most of the other segments.

The constraints relevant to producing a lattice of word hypotheses are sequential constraints, durational constraints, path constraints, and broad allophonic constraints. The lexical access component first applies sequential constraints to produce word hypotheses with associated time endpoints. Because many word candidates are hypothesized, other constraints are useful in reducing the number of candidates to a manageable number for verification. Durational constraints dictate whether a segment must represent one or two phones. For example, when duration constraints dictate that a segment represents only one phone, then all paths which require the segment to represent two phones can be removed. Path constraints, which do not require any extra processing before they can be used, are applied next to rule out word candidates which would form an incomplete path. Broad allophonic constraints are used to remove words from the lattice with context requirements which are incompatible with all the previous or following words.

Verification uses knowledge about detailed acoustic features of speech and detailed phonetic characteristics. Each phone is scored based on how well the acoustic features of the unknown segment match the expected acoustic feature values dictated by the phonetic characteristics of the hypothesized phone.

Thus the model uses knowledge which can be derived robustly to constrain the task at each stage of processing. Broad phonetic constraints are applied first, followed by more directed use of detailed constraints. Although use of only low-level speech knowledge is addressed, higher level knowledge can be incorporated into the model in the verification component.

2.5 Chapter Summary

In this chapter, the background leading to a model for continuous speech recognition was developed. The main issues discussed were:

- Shipman and Zue showed that strong sequential constraints exist on the words in English at the broad phonetic level. Based on this result, they proposed an isolated word recognition model.

- Strong sequential constraints also exist in continuous digits; therefore, the Shipman and Zue model can be extended to continuous digits.

- Many different types of low-level speech knowledge may be applied to the recognition task.

- A model for continuous speech recognition based upon the use of broad phonetic sequential constraints was proposed. Other types of speech knowledge were also incorporated into the model.

Chapter 3

Broad Phonetic Classification and Lexical Access

Speech from the same speaker saying the same phrase is never identical, and speech from different speakers contains even greater differences. These interspeaker and intraspeaker differences in speech occur because natural speech is not a sequence of discrete units; it is a continuum of sounds. Speech sounds are a function of the jaw and tongue position, and the continuous movements of the articulators modify the vocal tract configuration to produce a time varying signal. Depending upon the speaking rate and the current configuration of the articulators, the "target" for the next sound is reached with different degrees of accuracy before movement begins toward the configuration of the following sound.

Utterances of the same sentence vary in rate, pronunciation of each word, and degree of coarticulation. Speaking rate is affected by factors such as a speaker's mood and to whom the speaker is talking (e g., child or adult). As the speech rate varies, the durations of different speech souuds vary nonlinearly; for example, as the speech rate becomes slower, vowel duration increases much more than stop burst duration. A speaker may pronounce the same phoneme using different realizations because of phonetic context, as in the example with /u/ in alveolar and labial

contexts, or for other reasons. For example, a person may sometimes say "eight" with a released /t/ and sometimes with an unreleased /t/. A speaker may even delete sounds, such as pronouncing "eight six" as [eʸsɪk˺ks], where the /t/ in "eight" is deleted, or insert sounds, such as pronouncing "three" as /θəri/ where a /ə/ has been inserted. In addition, different acoustic realizations of a phoneme may be used by different speakers. For example, speakers may pronounce an intervocalic /v/ as a canonical voiced weak fricative, an unvoiced weak fricative, or the /v/ may be so weak that it approaches silence.

A speech recognition system must handle these variables. The system must know about common factors for each sound and/or know about the causes of variation and use this knowledge in recognition. In our implementation of the continuous speech recognition model, variability in natural speech was handled at two levels. The broad phonetic classifier dealt with variability within a class of sounds, and the lexical access component dealt with segmentation errors due to variability in the broad class representation of a sound.

A broad phonetic representation must be at least as accurate as a detailed phonetic representation because a broad phonetic representation is a description that is embedded within a detailed phonetic description. Thus, less detailed distinctions need to be made to produce a broad phonetic representation than a detailed phonetic transcription. In a broad phonetic representation, many speech variabilities, such as whether an /u/ is fronted are of no significance. The phonetic classifier labels an /u/ as a vowel whether or not it is fronted. Thus by knowing what characterizes different broad classes of speech sounds, the broad phonetic classifier labels speech at a broad phonetic level much more accurately than it could label speech at a detailed phonetic level.

However, segmentation errors can still happen at the broad phonetic level and unanticipated acoustic realizations can still occur. For instance, incomplete closure may occur in a stop gap, resulting in a "noisy" stop gap (see Figure 3.1) which

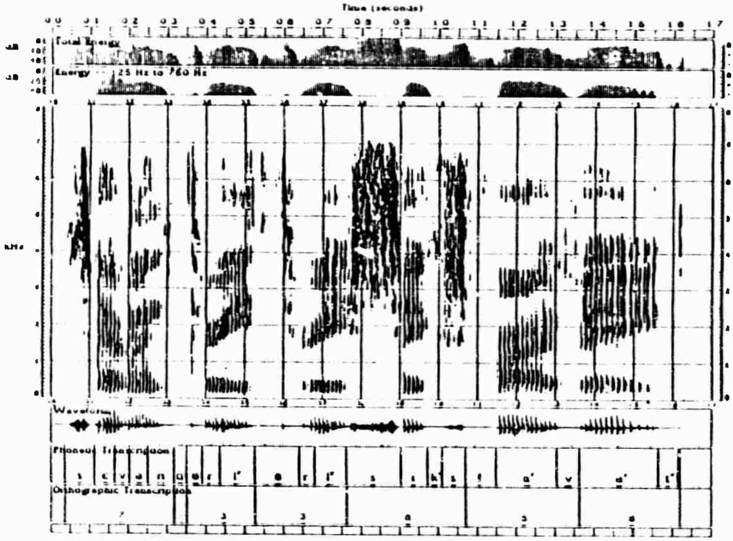


Figure 3.1: Noisy stop gap in "six" in digit string "733658"

may be labeled as a weak fricative. The lexical access component handles these segmentation errors due to speech variability using two types of knowledge: 1) how often a phoneme is mislabeled as another class and 2) how often a phoneme is labeled as a particular class given its context. An example of the first type of knowledge is how often a /k/ closure is labeled a "weak-fricative" instead of "silence" by the broad phonetic classifier. The second type of knowledge includes knowledge about when insertion or deletion errors occur. Examples of this type of knowledge include how frequently the /n/ and /s/ in the sequence /ns/ are both labeled as "strong-fricative" (a deletion) and how frequently an /n/ is labeled as a "sonorant" when preceded by an /s/ which was labeled a "strong-fricative" (a match). Together the broad phonetic classifier and lexical components take into account many of the

variabilities in speech to produce viable word candidates.

3.1 Broad Phonetic Classification

The broad phonetic classifier segments the speech signal and labels each segment as either silence or a broad phonetic class: strong fricative, weak fricative, short voiced obstruent, sonorant, or vowel. These classes were chosen because they could be robustly identified and correspond to different manners of articulation.

Classification is done by parameterizing the speech signal, extracting acoustic features from the parameters, and labeling segments based upon the acoustic features present. In an effort to minimize labeling errors, robust information as well as delayed binding was used in the classification approach. By allowing multiple segment labels until the final stage in processing, the classifier can use all the information learned from earlier stages in processing to make a final decision on labeling the segments.

3.1.1 Parameterization

Parameters were computed from speech digitized at 16 kHz and lowpass filtered at 6.4 kHz. The sampling rate was chosen to include the frequencies containing most of the speech information. Disagreement exists about the frequency range of speech: Hyde (1972) stated that speech "covers a frequency range of about 10 kHz"; in contrast, telephone speech has a passband of 300 Hz to 3300 Hz and is acceptably intelligible–however, this may be due to use of higher level constraints by listeners. In this study, the overriding consideration in choosing a sampling rate is the fact that Zue and his students can read spectrograms computed from speech sampled at 16 kHz and filtered at 6.4 kHz, indicating that enough information is present in the speech signal to be recognized when processed in this way. The information in the higher frequency range is important in identifying phonemes with energy concentrations in the higher frequencies (e.g., /s/). This is especially true in the analysis of female speech, which has higher natural frequencies due to the shorter average vocal tract length of females.

Acoustic parameters were defined for representing the speech signal in a compact format. Based on spectrographic examination of digit strings, candidate parameters which appeared to capture robustly the occurrence of significant events in a spectrogram were designed. The parameters in the final set were chosen for their usefulness in identifying and differentiating among different classes and for their robustness. The chosen parameters are energy in various frequency bands and zero crossing rate.

In most of the parameter computations, pre-emphasized speech was used since pre-emphasis compensates for the spectral tilt of the speech spectral envelope. This gives the higher frequencies, which are predominant in sounds such as /s/ and /θ/, approximately equal weight to the lower frequencies. The speech waveform was normalized before any computations so that the maximum of the sample values in each utterance is the same. The energies were calculated as log energy to minimize sensitivity to small variations when the energy values are large. To avoid rapid changes in value as a formant moved in or out of an energy band, tapered frequency windows were used to compute the energy within a frequency band from the DFT. The DFT's were computed every 5 msec using a 25.6 msec Hamming window. A sample tapered frequency window is shown Figure 3.2 and the corresponding frequency points used in the energy calculations are shown in Table 3.1.

A spectrogram of the digit string "6861994" is shown in Figure 3.3b, and corresponding parameters used in coarse acoustic analysis are shown in Figure 3.3a. Note that the low-frequency energy (energy 125-750 Hz) is highest in vowel and sonorant regions. This is because F_1, (and possibly a nasal formant) is present during the production of vowels and voiced sonorants. Thus low-frequency energy is a good indicator of voiced regions.

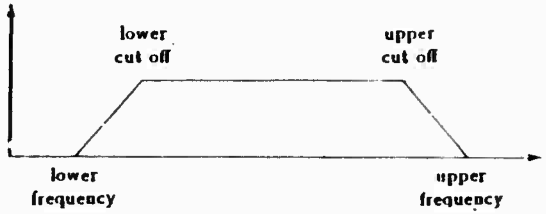


Figure 3.2: Sample Tapered Frequency Window

Table 3.1: Tapered Energies Used in Coarse Acoustic Analysis

Parameter	lower frequency	lower cut off	upper cut off	upper frequency
energy 125-750 Hz	0	125	750	900
energy 1000-2000 Hz	800	1000	2000	2200
energy 1000-3000 Hz	800	1000	3000	3200
energy 4500-7800 Hz	4300	4500	7800	8000

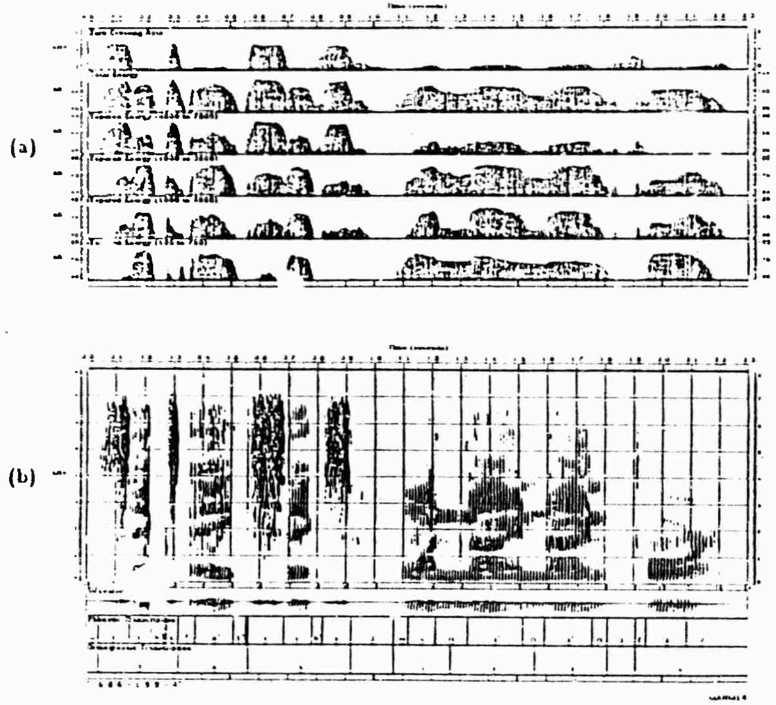


Figure 3.3: The Digit String "6861994" (a) Sample Parameters aud (b) Spectrogram

Intervocalic /n/'s have been observed to dip rapidly in the lower-mid-frequency energy (energy 1000-2000 Hz). The dip is due to a nasal zero in that frequency region aud the abruptness of the dip is due to the output quickly switching from the oral to nasal cavity. Note in Figure 3.3 the dip in lower-mid-frequency energy during the intervocalic /n/ of the "one nine" and "nine nine" sequences at around 1.25 and 1.55 sec, respectively. Other energy parameters show a slight dip during production of the intervocalic /n/, but the lower-mid-frequency energy dip is much more robust. This example stresses the importance in selecting appropriate parameters to characterize an acoustic feature.

Total energy is lowest during pauses and stop gaps (e.g., at 1.0 and .24 sec in Figure 3.3); it is usually lower in weak fricatives than strong fricatives (compare total energy in the /s/ at 0.10 and the /f/ at 1.92 sec). Thus total energy can be useful in indicating silence and in discriminating between weak and strong fricatives.

High-frequency energy (euergy 4500-7800 Hz) is largest in the presence of strong fricatives. The figures show that fricatives, especially /s/ and /z/, generally have more energy in the higher frequencies than vowels. This is because vowels have formauts beginuing at about 300 Hz, and each formant "tilts" the spectrum above it down by 12 dB/octave (Fant, 1970). In contrast, the lower frequency poles in fricatives are canceled by zeros; consequently the energy in fricatives is concentrated in the higher frequencies.

Strong vowel formants are usually observed in a spectrogram up to at least 3000 Hz, and even higher for front vowels. In many nonsonorants, like /v/, there is little euergy in this region. The mid-frequency energy parameter (energy 1000-3000 Hz) tries to capture the contrasting presence of mid-frequency energy in vowels and lack of it in consonauts. In some consonants there is significant energy in the mid-frequeucy range, such as a rounded /t/ in "two," but in these cases, low-frequeucy energy, which is usually weak during voiceless consonants like /t/, may be used as a secoudary cue to rule out the caudidate as a vowel.

The sero crossing rate is usually high during the production of fricatives (time 0.10 sec) and aspiration, because of the turbulence associated with the production of these sounds. The sero crossing rate is generally much lower in vowels, although the amount depends primarily upon the amount of high-frequency energy present in the speech of a speaker.

3.1.2 Acoustic Feature Extraction

During an utterance, parameters, such as the ones described above, exhibit salient features corresponding to speech sounds. For example, the sero crossing rate is high during voiceless fricatives. Algorithms for extracting a set of these *acoustic* features, as defined in Chapter 2, were designed. The feature set was composed of the descriptors ***high***, ***low***, ***dip***, and ***rapid transition***. ***High*** and ***low*** (high-low features) indicate the value of parameter in a region relative to values over the whole utterance and a set of standard value ranges. ***Dip*** indicates a region of lower value within a region classified as high. ***Rapid transition*** indicates that a region or dip has a rapid onset or offset. Not all parameter descriptors were computed for each parameter; instead only those descriptors which are robust indicators of a class or several classes of speech sounds were computed.

Broad classes of speech sounds have relatively stable characteristics during the middle portion of a segment. For example, vowels exhibit voicing. In contrast, the characteristics may change at different times during transitions between two broad classes of sounds. For example, in the transition from a vowel to a fricative, energy in the higher formants may weaken sooner than energy in the first formant. To avoid forcing a decision at each sample, which would result in less certain decisions in transition regions, the high-low descriptors label only robust *regions* where a parameter is relatively stable, and other regions are left unlabeled.

The high-low regions were found using an algorithm which depends on two thresholds, T1 and T2, to locate a region and then define the edges of a region. By

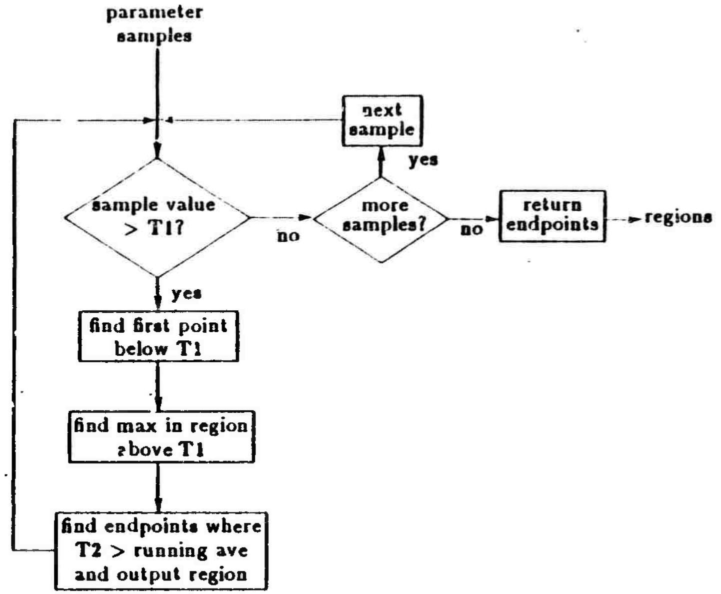

parameter samples
next sample
yes
sample value > T1?
no
more samples?
no
return endpoints
regions
yes
find first point below T1
find max in region above T1
find endpoints where T2 > running ave and output region

Figure 3.4: Flowchart for Finding "high" Regions

using two thresholds, robust regions, islands of reliability (Woods, 1981), can first be identified, and then anchoring from each robust region, the edges of the region can be extended. The high-low acoustic features were defined for each parameter by choosing different values of T1 and T2.

A flowchart for finding peaks using the high-low algorithm is shown in Figure 3.4. To locate high regions, points where a parameter value is greater than T1 are found first. T1 is dependent upon the minimum and maximum values observed in the utterance to be recognized and a "standard" set of values derived from a set of training utterances. T1 was defined as:

$$T1 = c_1(max_u - min_u) + min_u$$

where the max for the utterance, max_u is.

$$max_u = \begin{cases} max_{observed} & max_{observed} > max_{global} - r(max_{global} - min_{global}) \\ max_{global} & otherwise \end{cases}$$

and the min for the utterance, min_u, is:

$$min_u = \begin{cases} min_{observed} & min_{observed} < min_{global} + r(max_{global} - min_{global}) \\ min_{global} & otherwise \end{cases}$$

Max_u and min_u provide adjustment of threshold values for each utterance, allowing some adaptation to different speakers and/or environment. The constant r was chosen empirically to be .3 and is used to specify the range of values of $max_{observed}$ for which max_u is set to $max_{observed}$.

The condition on $max_{observed}$ and $min_{observed}$ insures that max_u and min_u are set to observed values for the utterance only if reasonable $max_{observed}$ and $min_{observed}$ values were c[illegible]ed. When the maximum(minimum) value of a parameter in the input utterance was within r of the "standard" maximum(minimum), the observed maximum(minimum) was considered reasonable and the sentence was assumed to contain at least one phone for which the parameter usually reaches the maximum(minimum) value. Conditionally adjusting the threshold in this way prevents errors such as lowering the threshold which defines regions of high zero crossing rate when no fricatives or stops are present in the utterance.

Once a region has been located, its edges are found using T2. The parameter is smoothed locally from the peak using a running average to minimize local perturbations:

$$s[i] = c_2 x[i] + (1 - c_2) s[i - 1]$$

where $s[i]$ is the value of the smoothed parameter i samples away from the peak (call the peak value max_{local}), $x[i]$ is the original parameter and c_2 is a constant chosen to be .25. The endpoints of a high region are defined to be the time when the value of the smoothed parameter has fallen to a predefined fraction (c_3) of the difference in value between the peak in the region and min_u. Thus T2 is defined as:

$$T2 = c_3(max_{local} - min_u) + min_u$$

The use of T2 allows the endpoints of a region to be set relative to the peak value of the region and independently of T1. Additionally, if the parameter is noisy and several alternating samples dip below T1, one region is found, rather than separate regions. *Low* regions are found using the same algorithm, but with complementary thresholds and comparators.

Dips for a parameter are found only in *high* regions using a simple dip detector. The slope is computed on the median smoothed parameter, and candidate dips are hypothesized at the point that the slope changes from negative to positive (call it P_1). From this point a local max in the smoothed parameter is found on each side. If the minimum dip depth, defined to be the minimum of the difference between the parameter value at each local max and P_1, is greater than 5 dB in the smoothed parameter, then the region between the two local maxima is defined to be a dip. Dips were computed on a 3-point median smoothed parameter and 7-point median smoothed parameter. The two different smoothers were used to capture short dips (3-point) and longer dips (7-point). The dips found on the 3-point and 7-point smoothed parameters were combined to form the list of dips for the parameter.

The transition rate is checked at edges of regions or dips associated with the energy from 1000 to 2000 Hz. For each edge, the maximum slope within 20 msec of the edge is found. If the maximum slope is greater than 1 dB/msec, then the transition is labeled "rapid."

To help moderate the effect of noise, redundant features can be used to indicate the presence of a phonetic class. Redundant features were in the same spirit in which Otten (1971) hypothesized that man utilizes redundancies in noisy situations.

A second algorithm for finding when the low-frequency energy is high was used. Rather than looking for regions where the parameter values are large, the second algorithm looks for edges using the fact that onsets in voiced sonorant regions

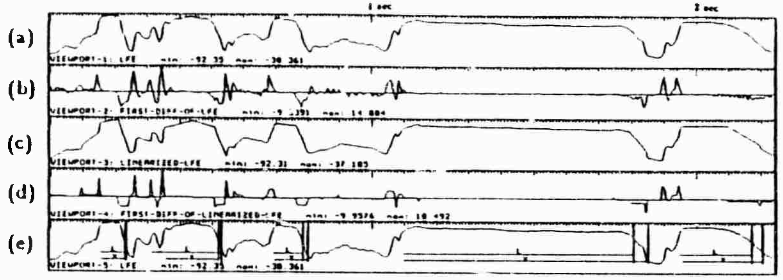


Figure 3.5: Low-Frequency Energy Characterizations of the Digit String "6861994" (a) low-frequency energy contour (b) first difference of the low-frequency energy contour (c) piecewise linear approximation to the low-frequency energy contour (d) first differeuce of piecewise linear approximation to the low-frequency energy contour (e) regions where low-frequency energy was found to be "high" using thresholding method "L" and edge detection method "v"

are generally characterized by a sharp rise in low-frequency energy and offsets are characterized by a more gradual decline in low-frequency energy. Thus onsets are detected first, and then an offset is found between each pair of onsets.

Figure 3.5 depicts parameters used iu edge detection. The edge detector method locates onsets and offsets (edges) in the low-frequcncy energy contour (Figure 3.5a) based upou the first differeuce of a piecewise linear approximation to low-frequency euergy (Figure 3.5d). A piecewise linear approximation (Pavlidis, 1974) was chosen for smoothing because it preserves the edges of the contour while smoothing small irregularities (compare (a) and (c)), resulting in a cleaner first difference of the linearized parameter (d). Because onsets in the low-frequency energy contour are sharper and can be more robustly detected than offsets, they are located where the

first difference of the linearized signal is large and the energy is above a minimal threshold. The energy threshold prevents false or early triggering, which happens when the spectral distribution of the energy in fricatives dips low. Between each pair of onsets, there must be an offset; the offset is found where the linearized first difference is most negative.

The output of the edge detector is shown in Figure 3.5e and regions between an onset and offset are indicated by a "v." Horizontal lines indicate the times the detector found the low-frequency energy to be high. The leftmost edge of a horizontal line marks the left endpoint and a vertical bar marks the right endpoint of the detected feature. The output of a threshold detector is also shown in Figure 3.5e; the regions in which the low-frequency energy was above a threshold are marked by an "L." Comparison of the two methods show close agreement in most cases.

When both detectors indicate that the low-frequency energy in a a region is high, then there is strong evidence that this is a voiced region. When there is disagreement between the two detectors, the low-frequency energy is not obviously high, but some acoustic event has occurred which causes the low-frequency energy to not be low. For example, this may be a stop with a release which extends into the low-frequency range. Thus by combining the output from both detectors, better decisions may be made.

Output from the feature detectors (rate information is not shown) for the utterance "6861994" is shown in Figure 3.6. The key to the symbols used in Figure 3.6 is given in Table 3.2. Note that most feature detectors turn on in a consistent manner: "h" is on during strong fricatives, "L", and "v" are on during voicing, and "D" is on during intervocalic nasals.

3.1.3 Broad Phonetic Labeling

Broad phonetic labeling uses a set of production rules to deduce possible broad classes from the chosen set of acoustic features. The hypothesized broad classes

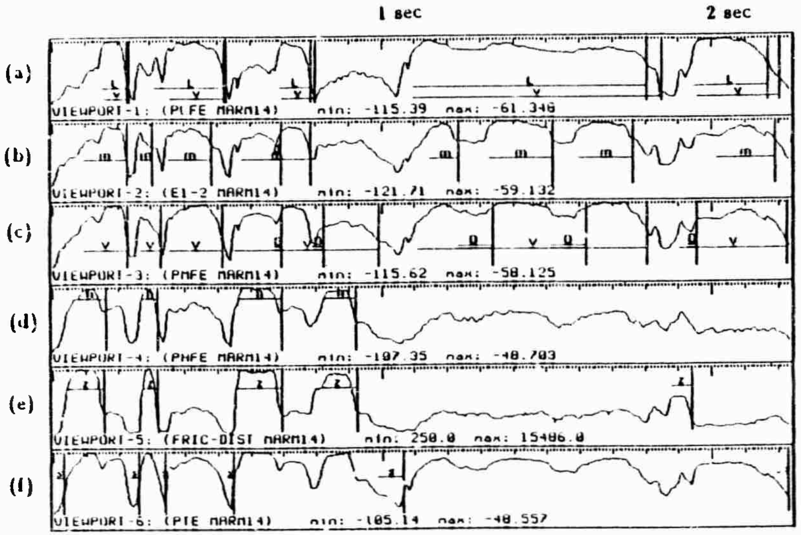


Figure 3.6: Feature Detectors for the Digit String "6861994" (a) energy 125-750 Hz is high using threshold "L" or edge detector method "v" (b) energy 1000-2000 Hz is high "m" or a dip occurs "d" (c) energy 1000-3000 Hz is high "V" or a dip occurs "D" (d) energy 4500-7800 Hz is high "h" (e) zero crossing rate is high "z" (f) total energy is low "s"

Table 3.2: Key to Figure 3.6

Symbol	Feature
h	high energy 4500-7800 Hz
z	high zero crossing rate
s	low total energy
m	high energy 1000-2000 Hz
d	dip in high energy 1000-2000
V	high energy 1000-3000 Hz
D	dip in high energy 1000-3000
l	high energy 125-750 Hz
v	high energy 125-750 Hz (edge-method)

form a segment lattice. When the production rules have deduced all candidate labels for each segment in the segment lattice, the computed values of the acoustic features are used to find the best label for each segment, thus producing a unique segmentation string. By first finding all possible labels, more directed analysis may be performed to reduce the segment lattice to a segmentation string by using knowledge of each competitor.

The production rules are applied in levels to produce the segment lattice. (See Appendix B for sample production rules.) By using multiple levels of rules, the knowledge gained by applying rules at lower levels can be used by higher level rules. Thus a rule can use contextual constraints requiring the presence of a preceding vowel if rules hypothesizing vowel-like segments have previously been applied.

The first set of 12 production rules hypothesizes each segment to be zero or more *phone-like* classes, based upon the presence or absence of combinations of non-conflicting robust acoustic features characterizing each segment. The acoustic features used are shown in the top of Table 3.3, and the phone-like classes used are shown in the center of Table 3.3.

The second set of 16 production rules hypothesizes *phone* classes from the phone-like classes using durational and contextual constraints to rule out some of the hypothesized phone-like classes. The duration of acoustic featu useful in classification; for example, an intervocalic, short, voiced obstruent is allowed a maximum duration of 55 msec and must be preceded and followed by a vowel. Contextual constraints are used to check that the context of a given label is correct. For example, an intervocalic, short voiced obstruent must be preceded and followed by a vowel. In addition, some speech sounds have slightly different realisations depending upon context. Intervocalic nasals are characterized by a sharp dip in lower-mid-frequency energy, but in non-intervocalic nasals, lower-mid-frequency energy may "fade out." To capture these differences within a class, a separate rule is used to describe each realization and its context. The phone classes used are shown in the lower portion of Table 3.3.

Figures 3.7a and b show the segment lattice produced by the classifier after the first and second sets of production rules are applied. The key to the symbols used in Figure 3.7 is given in Table 3.3. The segment from 0.23 to 0.26 sec is labeled as silence. It can be observed that silence segments are characterised by a low amount of total energy. Similarly, strong fricatives are characterized by a high sero crossing rate and a large amount of high-frequency energy, as in the segment from 0.82 to 0.92 sec, which is labeled as a strong fricative.

Each segment is not necessarily labeled as a broad phonetic class after the first two levels of production rules are applied. If there is conflicting information such that the cues are not robust enough to make a good decision (as in a transition), the segment is left unlabeled. Short unlabeled segments less than 40 msec in duration are handled first by arbitrarily splitting each evenly between the two adjacent segments on the assumption that they represent transitional regions. This produces a segment lattice without transitions and frees the lexical component from

Table 3.3: Key to Symbols Used in Figure 3.7

	Symbol	Description
acoustic features	h	high energy 4500-7800 Hs
	s	high sero crossing rate
	s	low total energy
	q	rapid transition in energy 1000-2000 Hs
	d	dip in high energy 1000-2000 Hs
	m	high energy 1000-2000 Hz
	D	dip in high energy 1000-3000 Hs
	V	high energy 1000-3000 Hs
	L	high energy 125-750 Hs
	v	high energy 125-750 Hs (edge-method)
phone-like classes	Fl	strong fricative like
	Wl	weak fricative like
	Sl	silence like
	Vl	short voiced obstruent like
	Rl	sonorant like
	VWl	vowel like
phone classes	F	strong fricative
	W	weak fricative aspiration
	S	silence
	V	short voiced obstruent
	R	sonorant
	VW	voiced sonorant, vowel

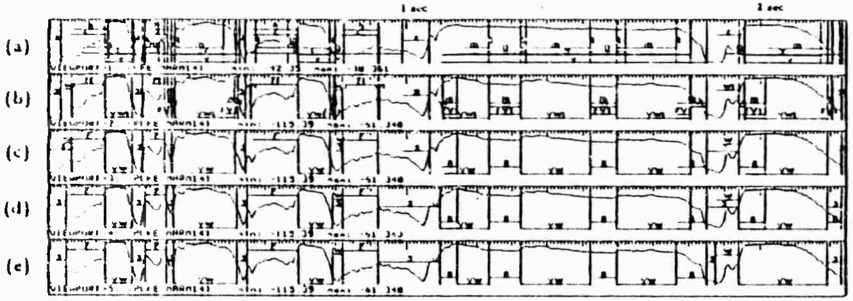


Figure 3.7: Steps in Broad Phonetic Labeling from Acoustic Features for the Digit String "6861994"

containing a subroutine for handling transition regions. In lexical access, boundary locations are irrelevant, only the order and approximate duration of the segments are important. In verification, the transition regions (boundaries) may be used to help identify a segment, but the central portion of the segment also contains much information which can be used for segment identification. Thus the philosophy was adopted that only the approximate location of boundaries need to be identified for recognition.

Unlabeled segments potentially match all broad classes and reflect no information. In these segments a consistent set of features was not present. A new feature set, called *not-features* are defined by relaxing the constraints on the high and low definitions. The not-features delimit regions where parameter values are "not high" or "not low" and are computed using the "high-low" algorithm. The set of features together with durational constraints are used to hypothesize possible label candidates and rule out definitely incorrect labels in the segments. In Figure 3.7c there

is an unlabeled segment beginning at time 0.92 sec. A corresponding segment lattice illustrating all labels after rules using "not-feature" values are applied is shown in Figure 3.7d; we can observe that the unlabeled segment beginning at time 0.92 sec was labeled as silence since all other possibilities have been ruled out by the not-features.

Once all broad phonetic labels are hypothesized for each segment, the segment lattice is reduced to a unique segmentation. Any segment which has been assigned more than one label is examined in more detail to determine the best label. Six acoustic features derived from the initial parameter set are used to capture parameter characteristics similar to the acoustic features used for initial segmentation. Since segment regions have been defined at this point in processing, computation can be performed over a specified region. Thus, rather than looking for a "high" region in an acoustic parameter, the maximum value of an acoustic parameter within a specified region is computed. The six acoustic features are: maximum pre-emphasized total energy in the center region, maximum pre-emphasized low-frequency energy in the center region, minimum total energy in the center region, minimum pre-emphasized lower-mid-frequency energy, maximum pre-emphasized mid-frequency energy, and maximum sero crossing rate. The center region, defined as the region from the quarter point to three-quarter point of the segment, was used to minimize transition effects on the computed values.

The label for each segment is chosen to be the label which is "most likely"; the likelihood of the label is computed based upon the distribution of the six acoustic features observed for each label and the value of each acoustic feature over the segment. In particular, the likelihood of label i, L_i, is computed as the product of the likelihood of label i versus label j, L_{ij}, over all j competitor labels:

$$L_i = \prod_{j \neq i} L_{ij} \tag{3.1}$$

The likelihood ratio of labels i and j is defined to be the product of the likelihood ratio of label i versus label j based on feature f, L_{ijf}, over the six acoustic features:

$$L_{ij} = \prod_f L_{ijf}$$

Thus the more likely each feature indicates that the label is label i, the more likely the segment is label i (rather than label j). Note from Equation 3.1, that the more likely a label is relative to each competitor label, the more likely it is that the segment is that particular label.

For each pair of competitor labels, i and j, L_{ijf} is computed from the value of the feature in the segment. The probability of a label given the observed feature value is estimated using k-nearest-neighbor estimation (Duda and Hart, 1973). The probability estimate for label l and feature f, P_{lf}, is used to compute the likelihood ratio between the labels i and j for feature f:

$$L_{ijf} = P_{if}/P_{jf}$$

Since a label is more likely the larger P_{lf} is, label l is more likely the larger L_{ijf} is. This algorithm was applied to the segment lattice to produce the final segmentation string shown in Figure 3.7e.

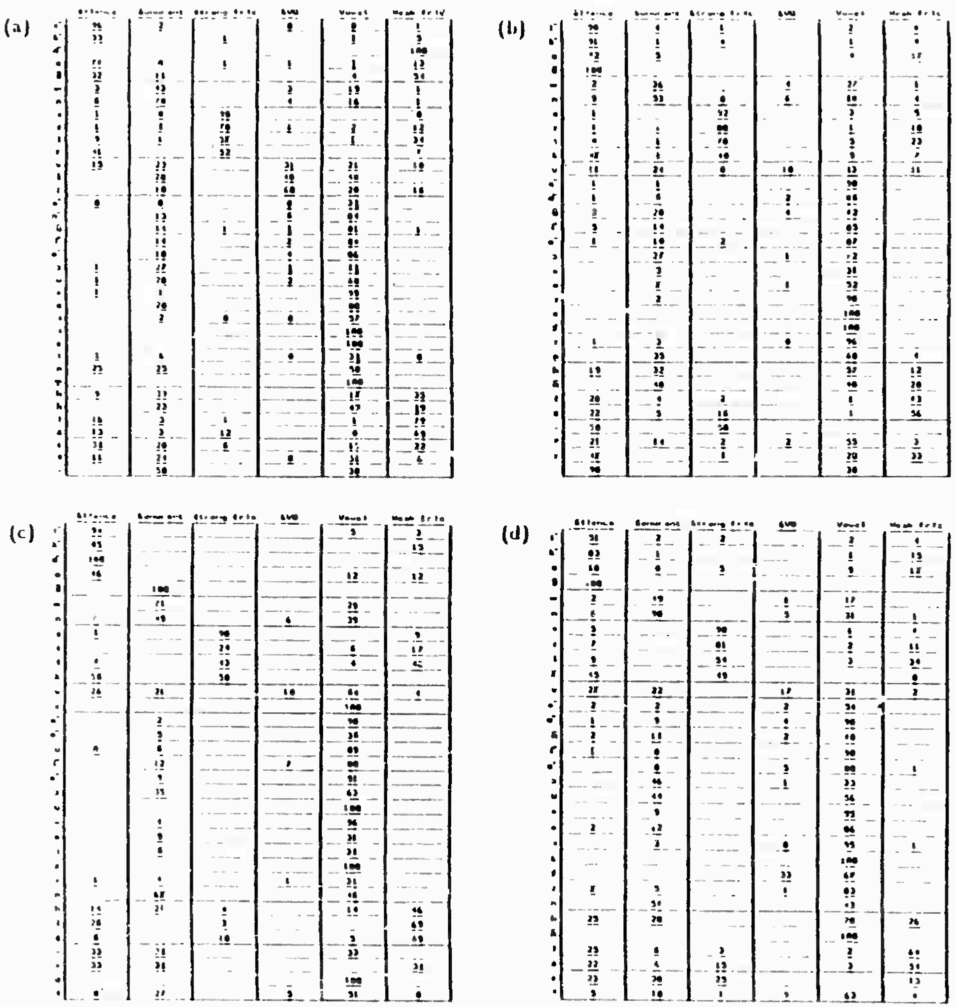
3.1.4 Evaluation

The output of the broad phonetic component was evaluated for insertions, deletions, substitutions, and matches when compared to a hand-labeled phonetic transcription (see Appendix A for a description of the phonetic labeling procedure). The broad class transcription of each utterance was derived by mapping each phone in the hand transcription to a broad class. The broad class hand transcription and automatic broad class transcription were aligued using a simple 50% overlap criterion: if segment A_i in string A covered over half the duration of segment B_j in string B, then segment A_i was mapped with segment B_j. An overlap criterion, rather than a string alignment was chosen because the time boundaries associated with

the segment regions are used later in verification. Since correct time boundaries are relevant, a string alignment is a less stringent criteria because matches are allowed even though the boundaries are shifted.

Insertions were defined as two adjacent segments from the automatic transcription mapping into one segment of the hand transcription. Deletions were defined as two or more adjacent segments of the hand transcription mapping into one segment of the automatic transcription. (See Appendix C for insertion and deletion errors.)

Substitutions were defined to occur any time the hand and automatic labels did not agree, based on the 50% overlap criterion, independent of whether or not an insertion or deletion occurred. This definition was used because the match statistics used by lexical access were computed this way. The confusion matrices of Figure 3.8 show the substitution errors and correct labels. Some errors were due to the limited set of labels used in hand transcription. For example, the stop gap for /k/ in the word "six" was always transcribed as [k⌝] if a demarcation was observed between the vowel and /s/, regardless of how noisy the closure was. Substituting "weak-fricative" for "silence" in this case is a reasonable error. Note that the bulk of the errors are reasonable, such as labeling /f/ (a weak fricative) as silence. Although /θ/ is also a weak fricative, in the digit lexicon /θ/ is always followed by an /r/. Since /r/ usually strengthens a preceding fricative, substituting strong fricative for weak fricative when a weak fricative is followed by an /r/ is also a reasonable error. A number of /n/'s were identified as a vowel. However, prevocalic /n/'s which may be a couple pitch periods in duration and therefore are not as salient as intervocalic /n/'s, are included in the statistics. Comparison of the performance for combinations of utterances and speakers reveals the performance to be similar. In cases where the labels differ, there are usually few samples, since these phones do not normally occur in digits. For example, voiced /h/ was sometimes used to mark aspiration at the end of sentence.



in addition to IPA symbols, additional symbols were used in the transcriptions: || for silence within a word, ⊔ for silence between words; + for noise; and - for voicebar.

Figure 3.8: Broad Phonetic Labeling Confusion Matrices for: (a) training utterance by training speakers (b) new utterances by training speakers (c) training utterances by new speakers (d) new utterances by new speakers

3.2 Lexical Access

Based upon the continuous speech recognition model, the lexical access component produces viable word hypotheses using its knowledge of the words in the lexicon, path constraints, how allophonic realizations of a phone are context-dependent, and reasonable durations for speech units. The component was implemented in two parts. First the word hypothesizer used sequential constraints to propose word candidates. Second, the word candidate pruner used durational, allophonic and path constraints to reduce the number of viable word candidates. The lexical access component could have been implemented so that checks on some of the pruning conditions are made as words are hypothesized. But by performing hypothesizing and pruning separately, the source of any errors may be more easily found. Since the purpose of developing the component was to study how speech knowledge could be applied to real speech for continuous speech recognition, the ability to quickly understand the source of error was important; consequently each constraint is applied as a separate step.

Lexical access produces a lattice of word candidates for the verifier. The application of speech constraints to prune the lattice produced by the word hypothesizer is important because the verifier can focus on reasonable candidates without performing computations on unreasonable word hypotheses, thus resulting in a more directed search.

3.2.1 Dictionary Representation

A word may be pronounced in multiple ways. The lexicon contains knowledge about allowable pronuuciations of each word and the context in which each word can occur. An average of two pronunciatious per word was used. Sentence-initial pronunciations allowing for a voice bar in "zero", a very weak initial /f/ in "five", and no initial closure in "two," were used. In addition, sentence-final pronunciations allowing for aspiration following the final /r/ in "four" and for deletion of the final closure in "eight" were used. Each phonetic pronunciation of a word is stored in an association list which is keyed by phonetic transcription. Associated with each pronunciation is a structure containing broad contextual information. This information is divided into four parts:

1. broad classes which must follow or cannot follow the current pronunciation. For example, [eʸɾ] requires a following vowel.

2. broad classes which must precede or cannot precede the current pronunciation. This information was not used because it was not applicable to the pronunciations used.

3. whether the first phone can geminate. The first phone in the pronunciation is not allowed to geminate when the first phone in the canonical pronunciation is deleted. For example, when the nasal murmur in the initial /n/ in "nine" is deleted, the /aʸ/ is not allowed to geminate with a preceding vowel, since intervocalic /n/'s are usually robust and should be found by the classifier.

4. whether the last phone can geminate. For example, a flap, as in [eʸɾ] is not allowed to geminate.

When a word is hypothesized, this information is kept with the word and accessed at the appropriate point in processing.

3.2.2 Hypothesizing Words

The word hypothesizer produces viable word candidates given the broad phonetic segmentation produced by the broad phonetic classifier and knowledge about the words which form the lexicon. In the ideal case where interspeaker and intraspeaker variations are minimal and the broad class segmentation is accurate, sequential constraints can be applied directly to the segmentation string. That is,

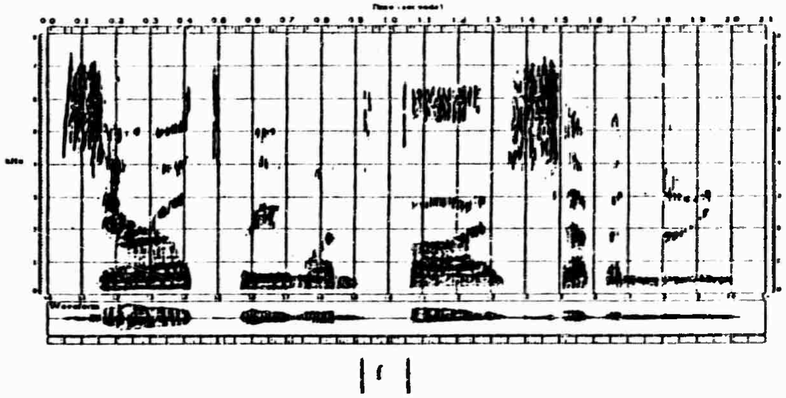


Figure 3.9: Spectrogram of the digit string "031579" with silence in /f/ in "five"

for each pronunciation of a word in the lexicon, which is represented by its phonetic and corresponding broad phonetic transcriptions, matches are found between a portion of the segmentation string and the broad phonetic representation of the pronunciation of a word.

Use of a segment lattice in place of the segmentation string can reduce the possibility of incorrect segmentation of real speech; however, a lattice handles speaker variations by enumeration. Different pronunciations produced by different speakers, must be included as alternate pronunciations, resulting in a very large lexicon. Furthermore, given an unanticipated realization of a word, the system will be unable to correctly propose the word. For example, a short period of silence in the middle of the /f/ in "five" may have been observed (see Figure 3.9), but the /f/ in "four" may have never been observed to be prounounced this way. If a speaker then says the /f/ in "four" with a short period of silence in the middle, the system should use the knowledge that it has seen /f/'s in other words pronounced this way. Thus the system should give the /f/ a good score, rather than commit a fatal error by assuming that the /f/ in "four" could never have some silence in the middle.

The use of a lattice was examined by performing lexical access on the segment lattice produced by the broad phonetic classifier immediately before the lattice is reduced to a segmentation string. The performance was evaluated on five new speakers by measuring how often the correct word was not among the word candidates. A word is considered a candidate when it overlaps in time with the spoken word by more than 50%. The correct digit was not one of the lexical candidates only 1% of the time. However, this measure did not insure that path constraints would be satisfied. For example, /θri/ is represented as "weak-fricative vowel" at the broad phonetic level. If the underlying /θ/ is classified as "weak-fricative silence weak-fricative" by the broad phonetic classifier, and this representation was not in the lexicon, then "three" would be a lexical candidate, but the "weak-fricative silence" portion of the segment lattice would not be associated with the three. Thus an alternate pronunciation of /θθθri/ would have to be added to the lexicon. It was found that a segment lattice did not provide enough flexibility and that the size of the lexicon had to be increased to accommodate new pronunciations.

Thus the word hypothesizer used knowledge about the characteristics of the segmentation strings produced by the front end. A scoring algorithm was developed to allow for some acoustic variations in a phone. Many alternate pronunciations needed with the straight matching method are unnecessary with this method because the system knows the types of errors that the broad phonetic classifier tends to make and uses that knowledge in scoring each word. For instance, the broad phonetic classifier may call /θ/ a weak fricative 60% of the time and a strong fricative 40% of the time. The system knows this and therefore when a /θ/ is called a strong fricative, the score is not reduced much. In contrast, if a /θ/ is never called a vowel, then the match of /θ/ to the broad class label "vowel" would be assigned a poor score. In addition to substitution errors, the algorithm also handles insertion and deletion errors by using transition probabilities. If the broad phonetic classifier

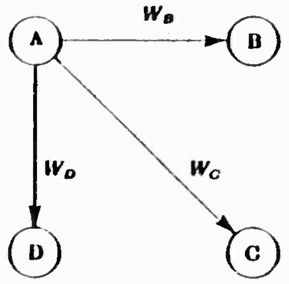


Figure 3.10: Paths Used in Dynamic Programming Algorithm

consistently misses prevocalic nasals, as in the word "nine," then the system will know that very often the /n/ as well as the /a^y/ is labeled as a vowel. This is reflected by high transition probability of matching /n/ to "vowel" and then matching /a^y/ to "vowel."

The general idea of the matching algorithm is to use knowledge about the characteristics of the broad phonetic classifier's output to assign a score reflecting how well a phonetic pronunciation matches a portion of the segmentation string. Each segment of the segmentation string is used as a beginning segment for matching each pronunciation. One or more end segments are associated with a beginning segment. Each end segment is chosen based on the length of the pronunciation's phonetic transcription; the end segments are iteratively chosen to be all segments within the range: $S_b + fixr(.5 * L)$ and $S_b + fixr(2 * L)$ where S_b is the beginning segment, L is the number of elements in the phonetic transcription, and $fixr$ is the operation of rounding to the nearest whole number.

A forward dynamic programming algorithm (Winston, 1984) was used to find the best match between the two sequences. The allowed paths from each node are illustrated in Figure 3.10. Simple 1:1 slope constraints are used, requiring the path to be monotonically non-decreasing in each direction. In contrast to the constraints used in dynamic time warping of the speech signal, many phonetic segments may map into a single label, as the /ɪ/, /r/ and /oʷ/ in "zero" maps into the label "vcwel," because the broad phonetic classifier has no knowledge for differentiating among these sounds. Figure 3.10 shows that three paths or transitions (to nodes B, C, and D) exit from a typical node A. The total accumulated score or "distance" to node D, d_D is computed as:

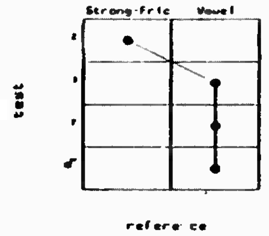
$$d_D = d_A + log[\Pr(p_D, l_D) \cdot W_D]$$

d_A is the total accumulated score to node A, $\Pr(p_D, l_D)$ represents the probability of the phone at node D, p_D, being labeled as the broad class label l_D. W_D is the probability of making a transition from node A to D, given that node A is the current state and nodes B,C, and D are states which may be entered from node A. W_D is computed as:

$$W_D = \frac{\Pr(p_A l_A \rightarrow p_D l_D)}{\Pr(p_A l_A \rightarrow p_B l_B) + \Pr(p_A l_A \rightarrow p_C l_C) + \Pr(p_A l_A \rightarrow p_D l_D)}$$

Thus W_D represents the probability of deleting a segment such that p_A and p_D map to l_A, which is the same as l_D. Similarly, W_B represents the probability of deleting a segment. The use of these weighting functions incorporates information about insertion and deletion probabilities into the score.

Figure 3.11 illustrates the alignment between the phonetic string /zɪroʷ/ and the broad phonetic representation "strong-fricative vowel." Probability scores used in the computation are shown on the right. The alignment score between each phone and broad phonetic class is shown under "match." The transition score from the previous node to the current node is shown under "weight." A weight is not shown for the first phone and broad class pair because a transition was not made. Not that an insertion or deletion occurs 1% of the time in the first transition only. The total accumulated score to a phone and label pair is shown under "total." The score assigned to a phonetic string is the total score of the best path. This score is normalized by the number of transitions and is shown as the final score in the figure.



reference	test	match	weight	total
STRONG-FRIC	z:	0.7754	1.0000	0.7754
VOWEL	I:	0.3714	0.9900	0.7457
VOWEL	r:	0.9236	1.0000	0.6608
VOWEL	o:	0.8414	1.0000	0.5795
				0.8725

Figure 3.11: Alignment of /zɪroʷ/ with "strong-fricative vowel"

By allowing each pronunciation of a word to begin at each segment with possible multiple end segments, many words can occur in a word lattice. For example, in a lattice represented by 20 broad phonetic segments, a word represented by four phones would be hypothesized 112 times. Although the number of candidates is large, this number is much smaller than the possible number of candidates in a frame-by-frame approach, such as dynamic time warping, where each word can begin at every frame. Furthermore, many of these hypotheses are poor matches, and some type of pruning can be applied to remove these poor matches.

Two word score distributions, on a $\log_e$ scale, are shown in each part of Figure 3.12. The distribution of correct word scores is indicated by the dashed line, and the distribution of the scores of all word hypotheses for a sample utterance are indicated by a solid line. Comparing the two distributions in each figure, we note that the log probability scores of the correct words are much closer to 0, or a probability of 1, than the bulk of the scores of all possible words. Note also that the distributions are similar for the new utterances spoken by training speakers and by new speakers, indicating the potential speaker independence of the approach.

A word score threshold can then be set such that all words with a score below the threshold are ruled out as a viable candidate. If a word is pruned as soon as

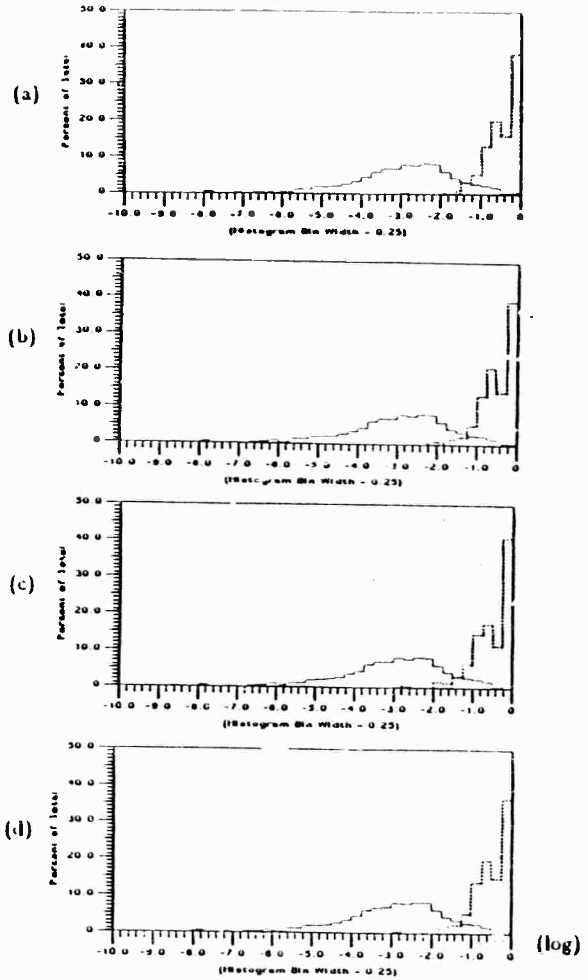


Figure 3.12: Histograms of Correct Word Scores vs All Word Candidate Scores: (a) training utterances by training speakers (b) new utterances by training speakers (c) training utterances by new speakers (d) new utterances by new speakers

the cumulative cost computed by the alignment algorithm passes the threshold, the amount of computation required for finding word candidates can be significantly reduced (by about an order of magnitude as measured from run times when the threshold is set at -1.5). Thus by setting the threshold at different values, the system can be biased towards less computation with more errors, or towards less errors and more computation, as desired.

Figure 3.13 illustrates the relationship between the amount of pruning achieved compared to the percentage of correct words pruned. We can observe the similarity of the curves and how the reduction of all word candidates is much more than the reduction of the correct word candidates for a given pruning threshold.

Figure 3.14 illustrates the relationship between the number of word candidates in the word lattice per word in the digit string as a function of the pruning threshold used in application of sequential constraints. Note that the number of word candidates increases sharply as the threshold is initially relaxed. The large number of word candidates when the threshold is very weak is due to the combinatorics of allowing words to begin and end at multiple segments, as explained earlier in this section.

3.2.3 Pruning the Word Lattice

Simple constraints based upon speech knowledge can be used to rule out very unlikely candidates. For example, if [e^{y}ɾ] ("eight" pronounced with a flapped /t/) is hypothesized, then the following word must begin with a vowel, since a /t/ is flapped only in the context of vowels. If none of the following hypothesized words begins with a vowel, then [e^{y}ɾ] can be ruled out as a viable word candidate.

Three types of constraints were applied following word hypothesis: path constraints, durational constraints, and allophonic constraints. The block diagram in Figure 3.15 illustrates when each constraint is applied. For example, durational constraints are applied first to rule out word candidates which depend on a seg-

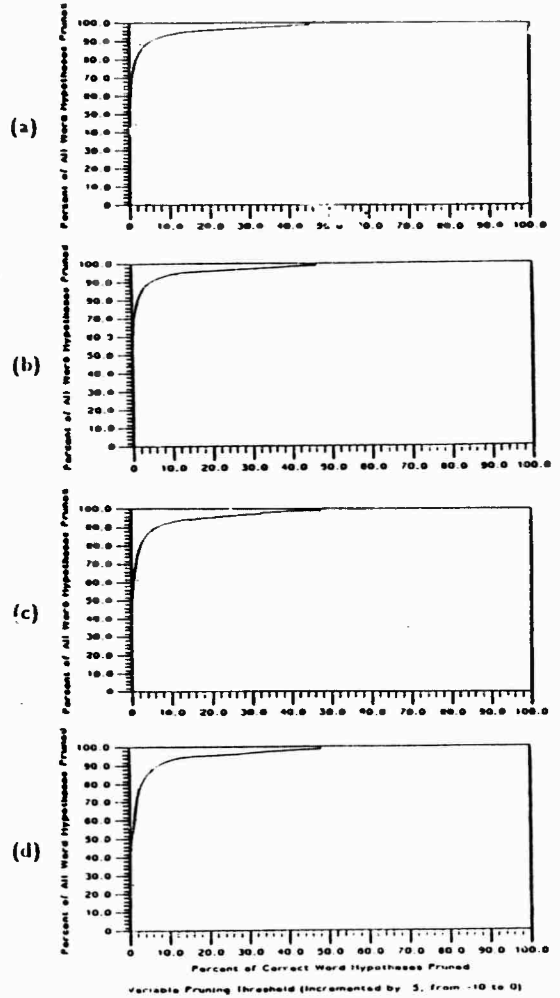


Figure 3.15: Pruning of All Word Hypotheses and Correct Word Hypotheses (a) training utterances by training speakers (b) new utterances by training speakers (c) training utterances by new speakers (d) new utterances by new speakers

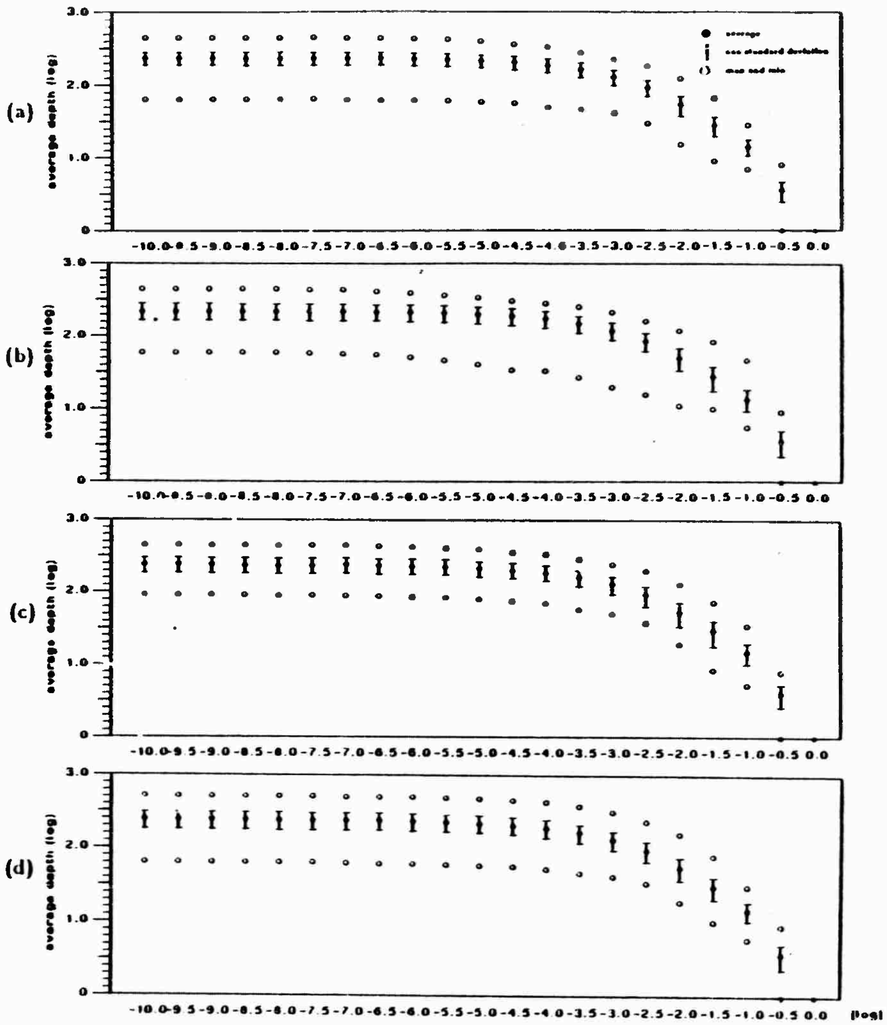


Figure 3.14: Lattice Depth vs Sequential Constraint Threshold (a) training utterances by training speakers (b) new utterances by training speakers (c) training utterances by new speakers (d) new utterances by new speakers

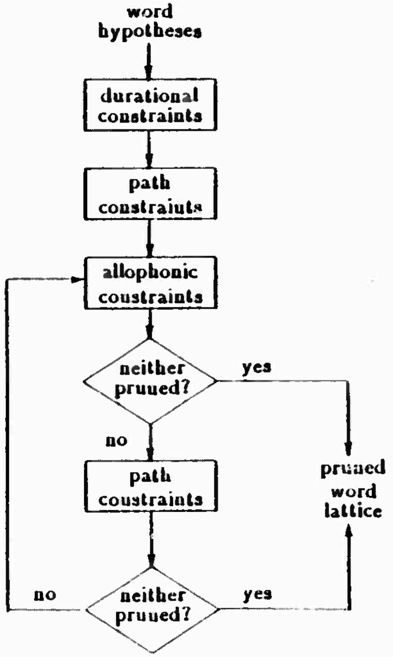


Figure 3.15: Application of Pruning Constraints

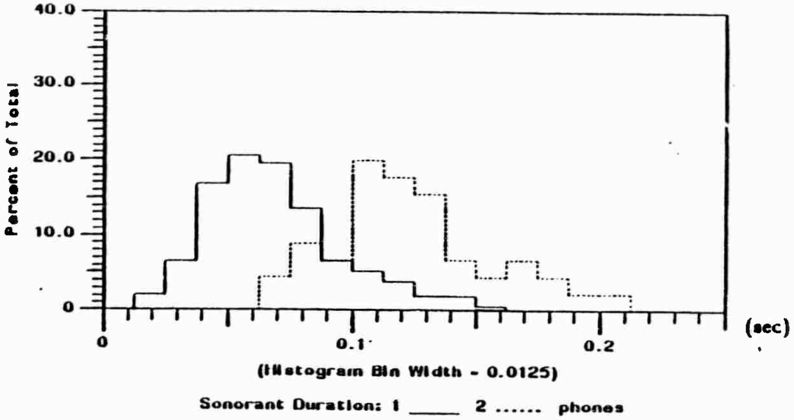


Figure 3.16: Distribution of sonorant duration for segments containing 1 or 2 phones

ment representing 1 phones when the durational training data indicates that 2 phones are represented by the segment. After duration constraints are applied, path constraints and broad allophonic constraints are applied iteratively until neither constraint prunes any word from the word lattice. Path and broad allophonic constraints are applied iteratively because removal of a word by one constraint may cause the conditions for the other constraint to remove a word(s) to be satisfied.

As an example, Figure 3.16 shows the distribution of sonorant segment durations in the training set (see Appendix A for a description of the training set). Table 3.4 shows the cutoff points for regions where only 1 phone, 1 or 2 phones, and only 2 phones occurred for the broad phonetic classes: "sonorant," "vowel" and "strong fricative."

Path constraints, as described in Section 2.3.2, require that each word in the lattice form part of a complete path. Words which do not have a legal "next word" and "previous word" are pruned from the lattice. The "next word" can be either a word which begins where the current word ends, the end of the sentence, or a word that has an initial phone which could acoustically geminate with the final phone of

Table 3.4: Cutoff Points of Segment Duration

Broad Class	Lower Cutoff (sec)	Upper Cutoff (sec)
Sonorant	.04	.17
Strong Fricative	.09	.15
Vowel	.09	.26

the current word; "previous word" is defined similarly. Acoustic gemination is the merging of two phones such that they are acoustically realised as one phone. This occurs when two adjacent phones are similar, such as the final /s/ in "six" and the initial /s/ in "sero." These two phones may merge such that they appear as one strong fricative.

Broad allophonic constraints, as described in Section 2.3.4, require that the context of a word be compatible at the broad phonetic level. The flapped /t/ in "eight" is an example of broad allophonic constraints. When none of the word hypotheses satisfy the contextual constraints of a word, the word can be removed as a viable word hypothesis.

Figure 3.17 illustrates the primary steps in lexical access. Each "box" in the figure represents the relative position of a label and does not convey any information about duration or rank order. The broad segmentation is shown in (a). Below in (b), all word candidates with scores better than the pruning threshold (chosen to be -0.75 for this example) are displayed in the word lattice. Application of duration constraints did not remove any word candidates in this example. The first application of path constraints removed the words marked with a sharp sign in (b), producing the word lattice shown in (c). All the words ending at the last vowel segment which could not geminate with a word beginning in the vowel segment, denoted by a "#" in (b) were removed because no word in the vocabulary is composed only of a sonorant. The first application of broad allophonic constraints removed

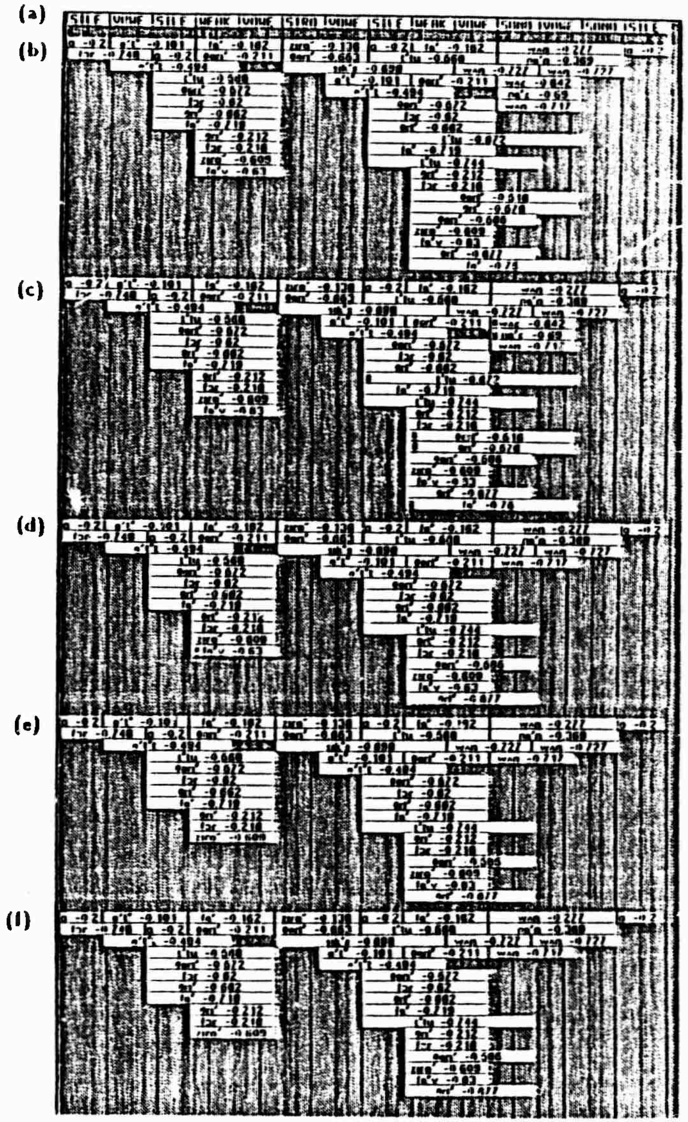


Figure 3.17: Pruning Word Candidates in Lexical Access

the /fa^{y}v/ marked with an asterisk in (d), producing the word lattice shown in (e). The /fa^{y}v/ was removed because in the context where a strong fricative follows /v/, the /v/ may be realized only as a vowel or fricative. Thus the pronunciation of "five" with the /v/ was removed, but the pronunciation with the /v/ deleted was kept. The final pruned lattice is shown in (f).

The pruning constraints were evaluated on the utterances in the training set by comparing the number of word candidates immediately after all words are hypothesized with the number of word candidates after pruning. Two sequential constraint thresholds were tested. The ratio of word candidates after pruning to word candidates before pruning was 0.69 at a sequential constraint threshold of -0.75 and the ratio was 0.76 at a sequential constraint threshold of -1.75. Thus application of durational, path, and allophonic constraints can reduce the number of word candidates even further. However, the constraint is not as strong when the sequential threshold is more lenient. With a weaker threshold (more negative), additional word candidates are allowed which can prevent the pruning constraints from being satisfied. Therefore, the pruning constraints need some prior reduction of word candidates before they can be applied effectively.

3.3 Chapter Summary

The main points addressed in this chapter are:

- Speech is highly variable. However, this variability is less evident in a broad phonetic representation. Thus many of the variabilities can be handled by representing speech at a broad phonetic level.

- The broad phonetic classifier segments and labels speech into broad phonetic classes using a set of production rules applied to coarse acoustic features. The broad acoustic features characterizing the speech signal are defined by identifying robust regions and then extending outward.

- A delayed binding approach is used to produce the final segmentation string; that is, the segment string is determined after all candidate labels are found.

- Even with a segment lattice, the lexicon becomes very large due to speech variations

- By using the characteristics of the broad phonetic classifier, sequential constraints can be effectively applied to real speech.

- Path constraints and broad allophonic constraints can be applied to reduce the number of candidates even further.

Chapter 4

Feature-Based Verification of Word Hypotheses

This chapter explores the use of detailed acoustic features for verification of word hypotheses. In our model, the input to the verifier is a lattice of word candidates produced by the lexical component. In this lattice, the most unlikely candidates have been removed. The task of the verifier is to select the best word or string of words from among the competing word candidates using a set of detailed acoustic features.

To verify the word candidates, each word hypothesis is represented as a sequence of phones, and each phone is characterized by a set of detailed acoustic features. The features were chosen to capture salient acoustic characteristics of speech sounds and to provide good discrimination among easily confused sounds in the digits. Two simple scoring algorithms were then used to demonstrate the feasibility of using these acoustic features for verification.

4.1 Characterization of Phones

Although many different recognition units could be used to score word hypotheses, phones were chosen as the basic recognition unit because they are suitable for defining many types of phonetic features and because of potential extendability to other recognition tasks. By representing each of the word hypotheses as a sequence of phones, features can be developed to characterize phones rather than whole words. Since the number of phones in a language is limited, a phone representation is potentially extendable.

Many of the defined features take advantage of the fact that a phone representation is used. A phone representation allows features to be defined over specific regions of time. In the middle of a phone region, characteristics of only one type of sound are present, while at the edges of the region, transitional information is available. Thus features can be defined to be minimally influenced by coarticulation or to specifically capture coarticulation effects. For example, the features for characterizing F_1 and F_2 were defined such that coarticulation effects on the computed values are minimal.

A phone representation also allows a wider variety of acoustic-phonetic constraints to be exploited more easily than a frame-by-frame representation. A phone can be characterized over its entire duration, such as by the maximum or average values of a set of parameters, rather than checking values on a frame-by-frame basis. In addition, many characteristics of speech sounds which are difficult to capture in a frame-by-frame approach are easily captured in a phone representation. For example, onset rate characterizes a phonetic event, such as a rapid stop release. This acoustic event may occur in one or two frames in a frame-by-frame approach and influences the overall score only in the one or two frames. In a phone representation, such speech events can be captured explicitly, and in an acoustic-phonetic approach, the information can be given equal weight with other feature information. Thus the onset rate can be given more import in the decision process in a phone representation than in a frame-by-frame approach.

A small set of acoustic features was carefully designed to capture salient characteristics of phones and detailed differences between similar phones. The large amount of information needed to represent a spectrogram can be reduced by defining and identifying acoustic features which robustly capture the occurrence of significant events in the spectrogram. Observations of phone characteristics in a spectrogram were used to select the initial parameters and features. The parameters were chosen to characterize important regions or events in the speech signal, and the features were chosen to quantify the parameters within a phone region.

4.1.1 Parameters

Three types of parameters, the energy within a frequency band, a measure of the location of spectral concentration, and the location of the offset of the first major peak in smoothed spectra, were used to characterise the speech signal. All parameters, except for the peak offset, were computed at a fixed rate. The information each parameter captures and how each parameter is computed are described below.

Energy

The energy in a frequency band, E, is computed by applying a frequency window, $\vec{W}(e^{j\omega})$, to the short-time spectra, $\vec{X}(e\ \ )$:

$$E = \log\left|\vec{X}(e^{j\omega}) \bullet \vec{W}(e^{j\omega})\right|$$

Trapezoidal shaped frequency windows were used to minimize sharp changes in energy as a formant moved in or out of the frequency band. The energy was computed as log energy to minimize the sensitivity to small variations in value when the value is large.

A Hamming window, h[n], was used to calculate the short-time spectra:

$$\vec{X}(e^{j\omega}) = \sum_{k=-\infty}^{\infty} x[k]h[n-k]e^{-j\omega k}$$

Unless stated otherwise, the duration of the Hamming window was 25.6 msec. By varying the width of the Hamming window and the default update rate of 5 msec, the sensitivity of the parameter to energy changes in the speech signal can be modified.

Spectral Concentration

One of the primary characteristics of voiced phones is the presence of formants; and associated with each formant is a concentration of energy. When two formants are close in frequency, these energy concentrations may merge so that accurate tracking of formant frequencies is difficult. However, in the identification of phones, particularly the limited number of phones in the digit vocabulary, gross characteristics of energy location may be sufficient.

Spectral weighting windows were used to characterize the location of energy concentrations associated with the formants of speech. The location of spectral concentration, S, was computed by applying a weighting window to the spectrum:

$$S = \frac{\vec{X} \bullet \vec{W}}{\sum_i |x_i w_i|} = \frac{\sum_i x_i w_i}{\sum_i |x_i w_i|} \qquad (4.1)$$

where the magnitude-squared spectrum, $\vec{X}$, and weighting window, $\vec{W}$, represent vectors normalized by the mean value of the vector elements. The vectors are normalized by the mean value of the vector elements to remove bias in the computed value. Additional normalization by the *magnitude* of each pair of vector elements was motivated by the idea that when x_i and w_i are similar, they will add constructively as $|x_i w_i|$, but when x_i and w_i are different, they will add destructively relative to $|x_i w_i|$. Thus if $\vec{X}$ and $\vec{W}$ are very similar, S will be close to 1; if $\vec{X}$ and $\vec{W}$ are very dissimilar, S will be close to -1.

The weighting window can be any real function over the frequency range; it may be thought of as a generalized form of the center of mass or first moment, in which the weighting function is linear with frequency. By specifying the sensitivity of the window to different regions of the spectrum, the weighting function can be tailored

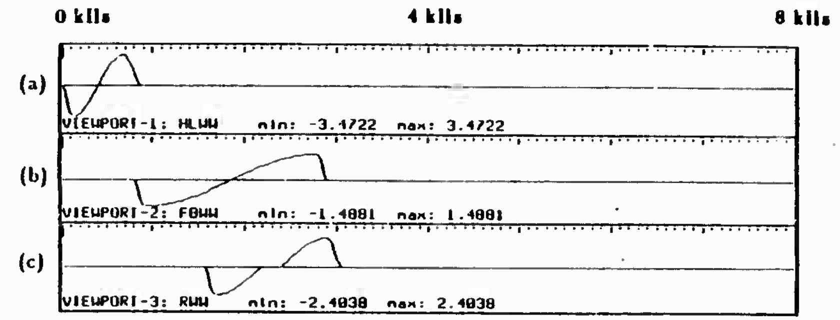


Figure 4.1: Weighting Windows: (a) High/Low (b) Front/Back (c) /r/

to capture particular spectral features. The weighting windows were tailored to be least sensitive near the edges of the frequency region in order to minimize the effect of formant motion at the edge of the windowed region. This was accomplished by defining weighting windowe with a slope of sero at the outer edges.

Figure 4.1 illustrates the three weighting windows used. These window positively weight the higher frequencies and negatively weight the lower frequencies. Thus the lower in frequency the energy is concentrated within the band, the lower the value of the spectral concentration. The first weighting window captures the position of energy concentration in the frequency range below 900 Hz. In vowels, this "high/low-spectral-concentration" parameter[1] roughly corresponds to the position of the first formant; and in nasal consonants, the value of this parameter is a function of the position of the nasal formant, nasal zero, and the first formant. The second weighting window provides information on the location of energy concentration in the frequency range from 900 Hz to 2800 Hz. This "front/back-spectral-concentration" parameter roughly corresponds to the position of the second formant in vowels and glides. The third weighting window attempts to capture when the third formant lowers in frequency for an /r/. During the production of an /r/, the third formant characteristically drops to approximately 2000 Hz. The exact frequency depends upon a number of factors, including the speaker and speaking rate. The amount of lowering also depends on whether the /r/ is prevocalic, postvocalic, or intervocalic; all three types of /r/ occur in the digit vocabulary. The "/r/-spectral-concentration" parameter looks for the presence of energy around 2000 Hz. Since the exact frequency to which F_3 dips is speaker-dependent, a "null" region was designed into the window, giving less import to exactly how low F_3 dips.

The high/low-, front/back-, and /r/-spectral-concentration parameters were designed to provide information about F_1 and F_2 and information about when F_3 lowers for an /r/. Sample output for the three weighting windows are shown in Figure 4.2. The /r/-spectral-concentration parameter approaches -1 when an /r/ is present. The values of the high/low- and front/back-spectral-concentration parameter usually correspond with F_1 and F_2, although the parameter values appear to change sharply. This is due to the shape of the window, which is more sensitive in the transition regions. In addition, note that the value of the high/low-spectral-concentration parameter at the end of the /w/ is larger than expected because of influence by F_2.

Offset of First Peak in Smoothed Spectra

The location of the offset of the first major peak in the shape of the spectrum was computed by finding the upper edge of the first peak in a cepstrally smoothed spectrum. Cepstrally smoothed spectra (Oppenheim and Schafer, 1968) were computed by applying a quefrency lifter which is constant the first 0.7 msec and cosine tapered the next 1.0 msec to the cepstrum of a 1024-point DFT. The DFT spectra were computed for the frequency range from 0 to 8 kHz. The extreme smoothing of the low-pass window produces a spectrum in which pitch harmonics are removed and only gross characteristics are evident. This type of spectral representation

[1] Some of the names given the parameters, features, and terms used in this chapter were chosen for conciseness at the sacrifice of accuracy. For example, the high/low-spectral-concentration does not always correspond with whether or not the vowel is high since this parameter may be influenced by F_2 when it is low in frequency

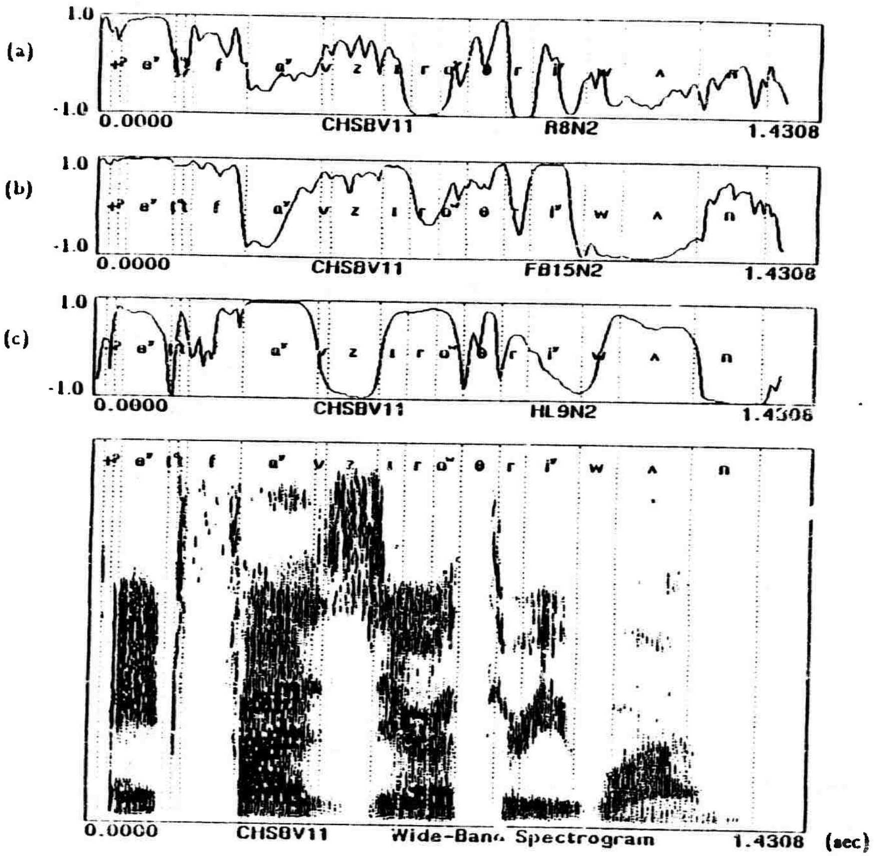


Figure 4.2: Sample Spectral Concentration Output for the Digit String "85031" (a) /r/-spectral-concentration (b) front/back-spectral-concentration (c) high/low-spectral-concentration

simplifies characterization of general spectral shape.

To locate the offset of the first major spectral peak, the smoothed spectrum was searched upward in frequency from the first peak until a value 12 dB down from the peak value was found. If a larger peak was encountered during the search, the peak value was set to the value of the larger peak. This parameter was computed for only one point per phone, partially because of the large amount of computation required to produce the smoothed spectrum. But unlike DFT spectra, this representation is relatively stable from sample to sample so that a high update rate appeared to be unnecessary.

4.1.2 Features

The values of a parameter within a phone region were converted to one feature value which characterizes some aspect of the parameter within the region. Nine acoustic features were defined for discriminating the digit phones. The nine features roughly describe the first three formants, the presence of a low frequency nasal pole, the onset rate of phones, the upper frequency of the first major concentration of energy, and changes in the noise source energy over the duration of a phone. A short name by which the feature will be referred to in later sections is given below for each feature. A more precise name and a description of the computation of each feature then follow. The nine features are:

1. **F_1-Normalized-Position:** This feature is the average value of the high/low-spectral-concentration over the middle 50% of the phone region:

$$\frac{\sum_{t=t_{.25}}^{t_{.75}} S_{F_1}(t)}{t_{.75} - t_{.25}}$$

where $S_{F_1}(t)$ is the high/low-spectral-concentration at time t, $t_{.25} = .25d_p + t_b$, $t_{.75} = .75d_p + t_b$, d_p is the duration of the phone in number of samples, and t_b is the sample at the time the phone begins. This feature is most useful for identifying vocalic phones and usually indicates whether a phone is more

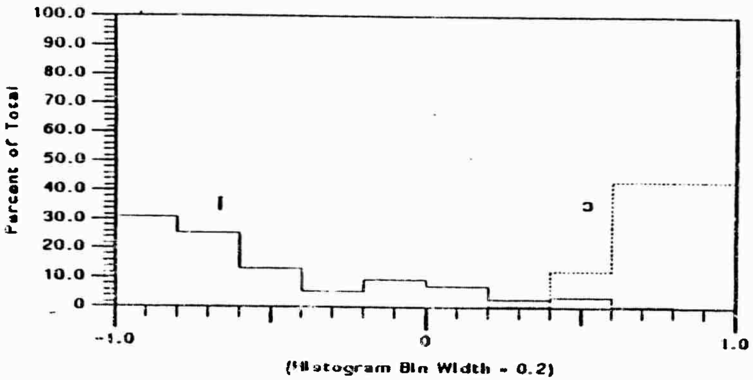


Figure 4.3: Distribution of F_1-Normalized-Position for /i/ and /ɔ/

like a high or low vowel. Figure 4.3 shows the distribution of F_1-Normalized-Position training ues for /i/ and /ɔ/. The values[2] for /i/ are generally lower than for /ɔ/ since /i/ generally has a lower F_1 than /ɔ/.

2. F_1-Movement: The movement of energy in the F_1 region is approximated by the slope of the best linear fit to the high/low-spectral-concentration over the middle 80% of the phone region:

$$\frac{N\sum_{t=t_{.1}}^{t_{.9}} t\,S_{F_1}(t) - \left(\sum_{t=t_{.1}}^{t_{.9}} t\right)\left(\sum_{t=t_{.1}}^{t_{.9}} S_{F_1}(t)\right)}{N\sum_{t=t_{.1}}^{t_{.9}} S_{F_1}(t)^2 - \left(\sum_{t=t_{.1}}^{t_{.9}} S_{F_1}(t)\right)^2}$$

where N is the number of samples from $t_{.1}$ to $t_{.9}$. This feature indicates how the energy below 1 kHz has shifted in frequency over the duration of a phone and is useful in discriminating between phones in which the average formant values are similar, but the amount of formant motion is different. Figure 4.4 shows the distribution of F_1-Movement training values for /aʸ/ and /ɔ/. The slope of /aʸ/ is more negative than the slope of /ɔ/ since F_1 falls in /aʸ/, but is approximately constant in /ɔ/.

[2] There are no units because spectral concentration parameters are normalized

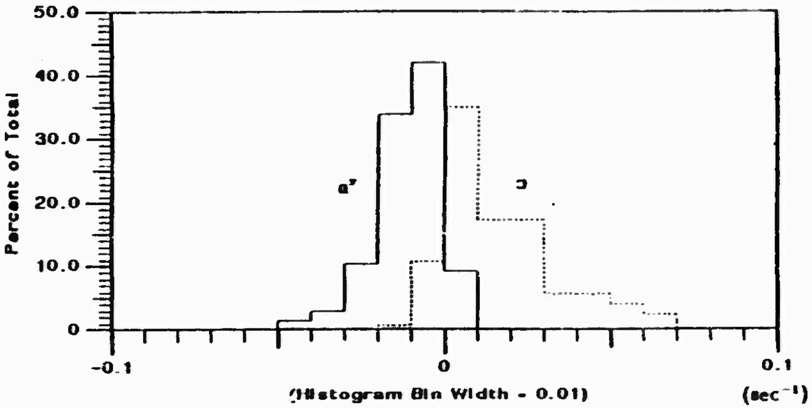


Figure 4.4: Distribution of F_1-Movement for /aʸ/ and /ɔ/

3. F_2 Normalised Position: This feature is the average value of the front/back-spectral-concentration parameter over the middle 50% of the ph ne region:

$$\frac{\sum_{t=t_{.25}}^{t_{.75}} S_{F_2}(t)}{t_{.75} - t_{.25}}$$

where $S_{F_2}(t)$ is the front/back-spectral-concentration at time t. This feature is most useful for identification of vocalic phones and indicates whether a phone is more like a front vowel or a back vowel. Figure 4.5 shows the distribution of F_2-Normalized-Position training values for /i/ and /ɔ/. The values for /i/ are generally higher than for /ɔ/ since /i/ has a higher F_2 than /ɔ/.

4. F_2-Movement: The movement of energy in the F_2 region is approximated by the slope of the best linear fit to the front/back-spectral-concentration over the middle 80% of the phone region:

$$\frac{N\sum_{t=t_{.1}}^{t_{.9}} t\,S_{F_2}(t) - \left(\sum_{t=t_{.1}}^{t_{.9}} t\right)\left(\sum_{t=t_{.1}}^{t_{.9}} S_{F_2}(t)\right)}{N\sum_{t=t_{.1}}^{t_{.9}} S_{F_2}(t)^2 - \left(\sum_{t=t_{.1}}^{t_{.9}} S_{F_2}(t)\right)^2}$$

This feature indicates how the energy in the range of F_2 has shifted in frequency over the duration of a phone. As with F_1-Movement, this feature is

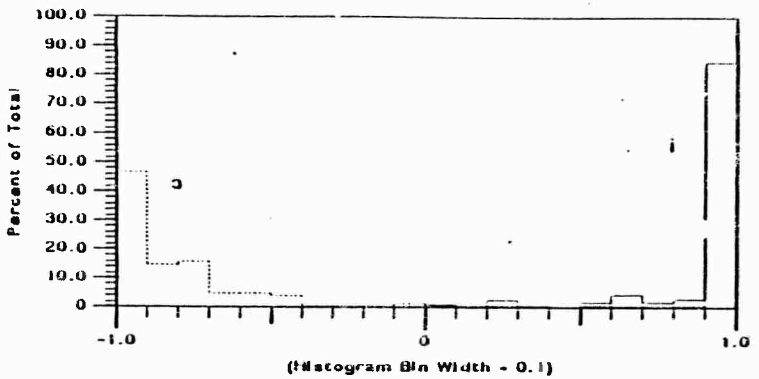


Figure 4.5: Distribution of F$_2$-Normalized-Position for /i/ and /ɔ/

useful in discriminating between phone pairs like /aʸ/ and /ɔ/. Figure 4.6 shows the distribution of F$_2$-Movement training values for /aʸ/ and /ɔ/. The slope of /aʸ/ is more positive than the slope of /ɔ/ since F$_2$ rises for /aʸ/ but is relatively constant for /ɔ/.

5. /r/-Possibility: This feature, which is useful in detecting /r/'s, is computed as the average value of the /r/-spectral-concentration over the middle 50% of the phone region:

$$\frac{\sum_{t=t_{25}}^{t_{75}} S_R(t)}{t_{75} - t_{25}}$$

where $S_R(t)$ is the /r/-spectral-concentration at time t. This feature indicates how much energy in the region of F$_3$ is below 2200 Hz, relative to energy above 2200 Hz. Figure 4.7 shows the distribution of /r/-Possibility training values for /r/ and /eʸ/. The values for /r/ are much lower than for /eʸ/ due to the lowering of F$_3$. The distribution includes pre-, post-, and intervocalic /r/'s. Better results should be obtained by computing the average value of /r/-spectral-concentration over different regions for the different allophones of /r/. That is, the feature for prevocalic /r/'s should be computed during the

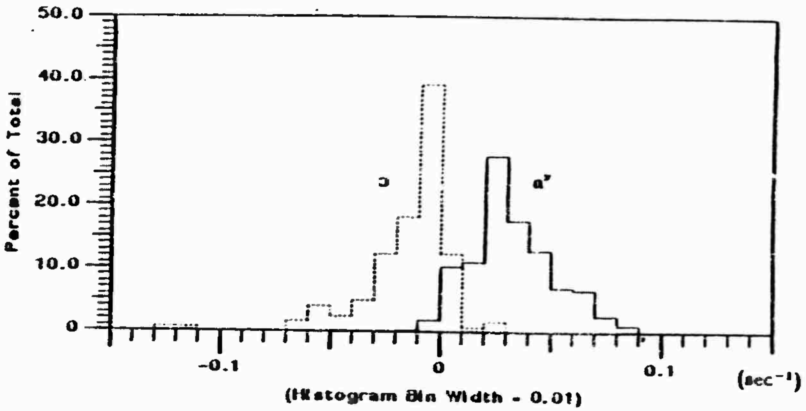


Figure 4.6: Distribution of F$_2$-Movement for /aʸ/ and /ɔ/

beginning portion of the phone, and the feature for postvocalic /r/'s should be computed during the final portion of the phone.

6. Nasal-Possibility: This feature is the average value of the difference in (log) energy computed with a passband of 100 to 350 Hz and (log) energy computed with a passband of 350 to 850 Hz:

$$\frac{\sum_{t=t_{25}}^{t_{75}} E_{100-350}(t) - E_{350-850}(t)}{t_{.75} - t_{.25}}$$

In each energy computation, 50 Hz tapers on the trapezoidal frequency window were used. This feature captures the presence of the low resonance around 300 Hz which is characteristic of nasal murmurs (Fujimura, 1962). Figure 4.8 shows the distribution of Nasal-Possibility training values for /n/ and /ɔ/. The values for nasal consonants are generally larger than for non-nasals due to the presence of energy from the low resonance in the lower band.

7. Onset-Rate: This feature measures the maximum change in energy from 1000 Hz to 7000 Hz within 20 msec of the beginning of a phone:

$$\max_{i=t_b-4}^{t_b+4} |E_{1000-7000}(i) - E_{1000-7000}(i-2)|$$

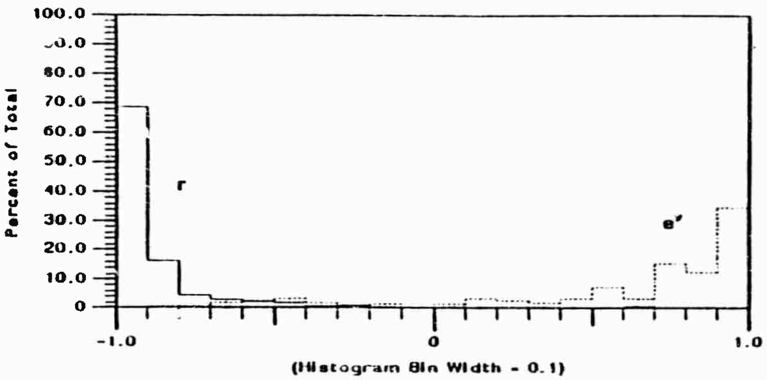


Figure 1.7: Distribution of /r/-Possibility for /r/ and /eʸ/

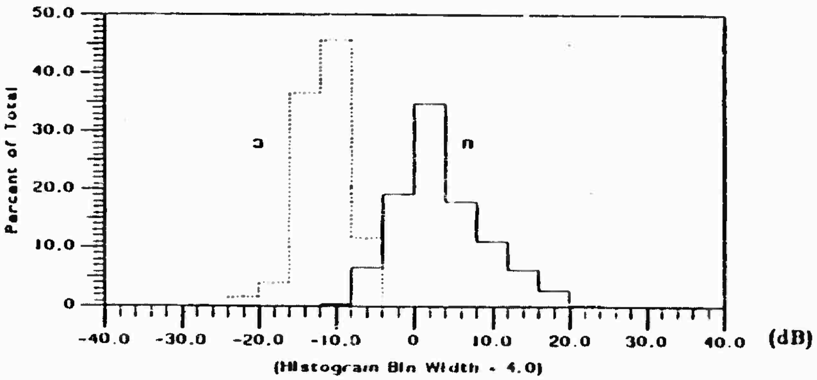


Figure 4.8: Distribution of Nasal-Possibility for /n/ and /ɔ/

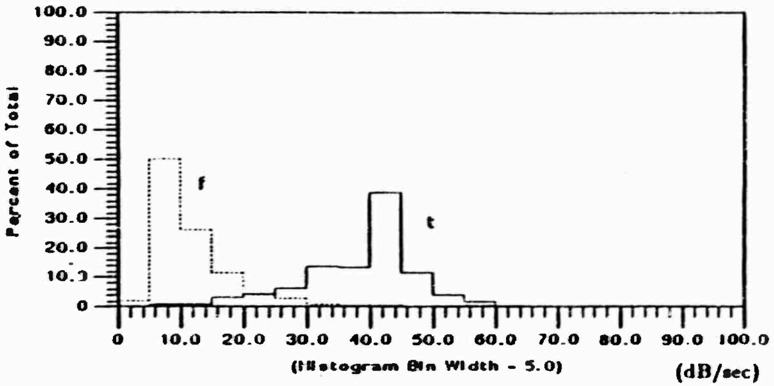


Figure 4.9: Distribution of Onset-Rate for [t] and [f]

where i is the sample number, t_b is the sample at the time the phone begins, and d is the number of samples in 20 msec. In order to capture rapid transitions, $E_{1000-7000}$ was computed every msec from the short time Fourier transform using a Hamming window width of 2 msec. This feature is particularly useful for discriminating stops from fricatives because stop onsets are generally much more rapid than fricative onsets. Figure 4.9 shows the distribution of Onset-Rate training values for [t] and [f]. The Onset-Rate is larger for the release of [t] than for [f] as expected.

8. **Spectral-Offset-Location:** This feature indicates the location of the first major spectral dip in the cepstrally smoothed spectrum 30% through the duration of the phone. The time at which this feature is computed was chosen empirically and was motivated by the task for which the feature was designed. This feature was initially designed to discriminate between the /aʸ/ in "five" and /ɔ/ in "four." The spectrogram in Figure 4 10 contains the word sequence "five four," and the location of the /aʸ/ and /ɔ/ are indicated below. Comparing the /ɔ/ in "four" with the /aʸ/ in "five," we note that the strik-

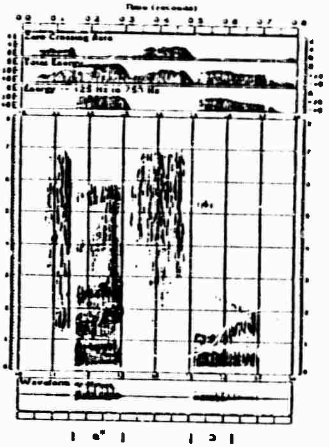


Figure 4.10: Spectrogram of "five four"

ing differences are in the location of F_2 and lack of energy between F_2 and F_3. This feature tries to capture the location of the upper edge of F_2 to differentiate between the two sounds. Figure 4.11 shows the distribution of Spectral-Offset-Location training values for /a^y/ and /ɔ/.

9. **High-Frequency-Energy-Change:** The change in high frequency energy is computed as the slope of the best linear fit to the energy in the 4500-7800 Hz band over the middle 80% of the phone region:

$$\frac{N\sum_{t=t_1}^{t_n} t\,H(t) - \left(\sum_{t=t_1}^{t_n} t\right)\left(\sum_{t=t_1}^{t_n} H(t)\right)}{N\sum_{t=t_1}^{t_n} H(t)^2 - \left(\sum_{t=t_1}^{t_n} H(t)\right)^2}$$

where H(t) is the value of high frequency energy at time t. This parameter is intended to help differentiate between fricatives and stop releases. Fricatives are relatively stable over their duration; in contrast, unvoiced stop releases generally have a strong onset followed by aspiration which weakens over the duration of the phone. Thus the slope is expected to be more negative for a stop release, such as |t|, than for a fricative such as |s|. Figure 4.12 shows

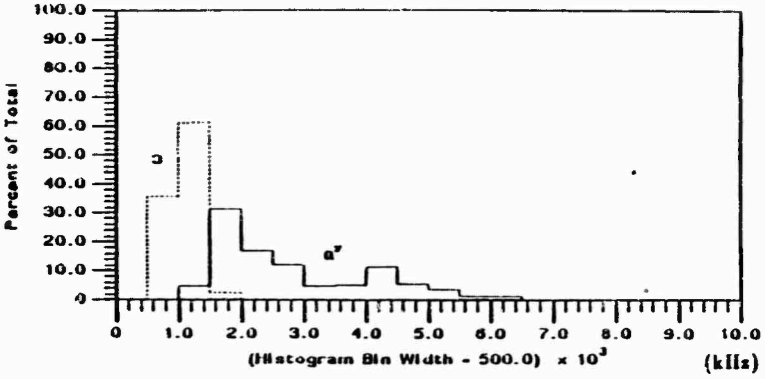


Figure 4.11: Distribution of Spectral-Offset-Location for /a^y/ and /ɔ/

the distribution of High-Frequency-Energy-Change training values for |s| and the release of |t|. The values for |s| center around 0, since fricatives generally do not weaken; the values for |t| are generally negative since high frequency energy usually decreases after the release in a stop.

4.2 Scoring Word Hypotheses

The task of assigning a score to each phone in a word lattice may be approached as a discrimination and/or identification problem. When viewed as a discrimination task, a binary discrimination is performed between each pair of competitor phones. The results are then used to assign a score to each phone indicating how good the hypothesized phone is relative to its competitors. When viewed as an identification task, each phone is assigned a score of how good it is, independent of the values of the other phones.

Since the lexical component has already reduced the competitors to a small subset of words, discrimination between the remaining competitors should give better performance than trying to identify a phone from all possible phones. In general,

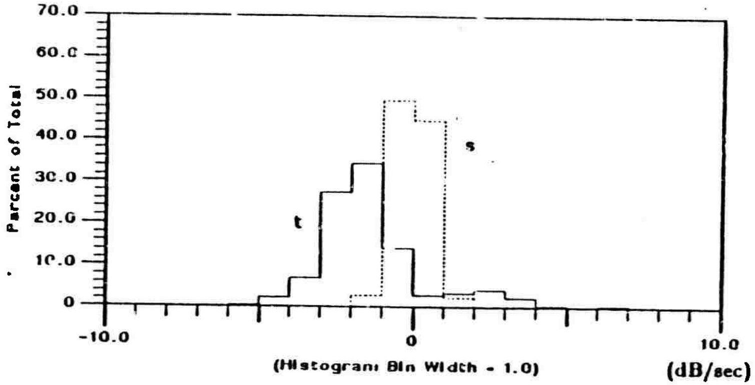


Figure 4.12: Distribution of High-Frequency-Energy-Change for |t| and |s|

it is easier to identify an object from a small group of objects by comparing it to each of the possible alternatives (discrimination) using knowledge about how they previously compared than it is to identify an object based only on knowledge of the characteristics of each object (identification). However, defining competitor phones in a meaningful way is difficult because phone boundaries are not always aligned. In contrast, identification does not require competitors to be defined. So although we expected discrimination to perform better, we explored both methods for verification. The same algorithm was used to compute word scores from the identification or discrimination scores, and the results from the two methods were compared. The identification and discrimination scores represent the score of a phone based upon information from particular detailed feature. These conditional phone scores were combined in this algorithm to produce phone scores, and the phone scores were combined to produce word scores.

Each approach required a priori knowledge of the distribution of the features used. This knowledge was provided from values computed on a set of training utterances described in Appendix A.

4.2.1 Phone Scores Based on Identification

In the identification task, observed values of the feature vector for a hypothesized phone and training values for the features are used to compute phone scores. The scores reflect the probability of a phone occurring given the observed feature values. In particular, the unnormalized score of phone i based on information from feature k, $S_u(p_i|f_k)$, was computed as:

$$S_u(p_i|f_k) = \Pr(p_i|f_k) = \frac{\Pr(f_k|p_i)\Pr(p_i)}{\Pr(f_k)}$$

where phone i is represented as p_i and feature k is represented as f_k. Assuming that each phone is equally likely, $\Pr(p_i)$ is a constant and can be ignored. Similarly, $\Pr(f_k)$ is the same for all phones being evaluated in the region and can also be ignored. Therefore, the score of phone i based upon information from feature k is proportional to $\Pr(f_k|p_i)$.

A k-nearest-neighbor estimate was used to compute $\Pr(f_k|p_i)$. Since the k-nearest-neighbor estimate is a function of the window width required to capture k samples surrounding a point, the estimate is a function of the range of each of the features. To permit information from different features to be combined, each estimate was normalized by the range, R, of the observed values for the feature over all phones. The normalized score of phone i based on information from feature k, $S(p_i|f_k)$, was thus computed as:

$$S(p_i|f_k) = \frac{\Pr(f_k|p_i)}{R}$$

4.2.2 Phone Scores Based on Discrimination

In the discrimination task, the score for each phone based upon a particular feature is a function of how well the phone compares to each of the competitor phones according to that feature. Discrimination of the phones in the word lattice produced by the lexical access component was expected to give better results than

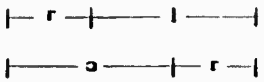


Figure 4.13: Alignment of /ri/ compared to /ɔr/

identification. This is because the recognition model used sequential constraints to remove unlikely word candidates from further consideration. The remaining word candidates are a small subset of all possible word candidates which can be discriminated using explicit knowledge of the limited number of competitors. Since each remaining phone candidate within a broad class segment matches the segment well and the phone candidates also match the initial features characterizing the broad class segment well, the remaining phones are similar to each other. Thus fine discriminations must be made in order to accurately score how well each phone is realized relative to the other phone candidates.

To compute a discrimination score, the phone being scored was first compared to each of its competitors. However, competitors can be defined in many ways. Ideally, competitors should cover the same region of an utterance. Since lexical access is performed at a broad phonetic level, one or more phones may map into one broad class, depending upon the hypothesized word. For example, /ɔr/ and /ri/ may both map to the broad class "vowel." Since /r/ in intrinsically shorter than the adjacent vowel, the boundary between /ɔ/ and /r/ will be after the boundary between /r/ and /i/, as illustrated in Figure 4.13

As seen in the previous example, the phones will not always line up, and a choice must be made as to whether or not all competitor phones should be forced to have the same endpoints. By requiring all competitors to have the same endpoints, either the boundaries of a phone will not be accurate, or the recognition unit will be composed of a variable number of phones. Consequently, separate acoustic features have to be developed which look for characteristics of a phone within a region. For

example, the vowel portion of the broad phonetic sequence "weak-fricative vowel silence" may map to /ri/, /rie*/, or /ɔr/. A separate feature for /r/ would then be required for each sequence. A problem that results from this approach is that as the vocabulary grows, the number of possible sequences grows. That is, a new word-initial(final) phone sequence may form new sequences with word-final(initial) phone sequences in the vocabulary when adjacent phones can geminate. In contrast, choosing the phone as the unit to be scored eliminates the need for separate feature detectors; only detectors for the phones in the vocabulary are needed.

Competitors were defined as any phone which overlaps in time with the phone being scored. The probability of phone i relative to each competitor phone j was computed using Bayes' Rule under the assumption that each phone is equally likely:

$$Pr_j(p_i|f_k) = \frac{\Pr(f_k|p_i)}{\Pr(f_k|p_i) + \Pr(f_k|p_j)}$$

As in identification, $\Pr(f_k|p_j)$ was computed using a k-nearest-neighbor estimate. Normalization by the feature value range was not performed since the score is computed as a ratio.

The score for phone i based on information from feature k, $S(p_i|f_k)$, was then computed as the average of how likely it is that phone i is the underlying phone relative to each competitor phone:

$$S(p_i|f_k) = \frac{\sum_{j=1}^{J} \Pr_j(p_i|f_k)}{J}$$

where J is the number of competitor phones. Thus the discrimination score is based on how phone i compares only to the competitor phones. This can be contrasted to the identification task where phone i is scored based upon how well it matches previous observations of the phone.

4.2.3 Computation of Phone and Word Scores

The score of a phone is a function of the set of scores for that phone conditionalized on different features. We can think of these scores as reflecting how strongly

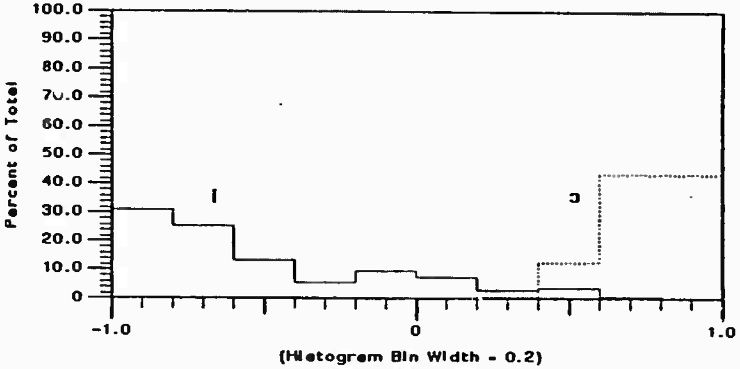


Figure 4.14: Percentage Histogram of /i/ and /ɔ/ Divided into 10 Bins

each feature indicates that the underlying phone is the hypothesized phone. Each feature provides input to the decision process, weighted by the quality of that feature in identifying the phone.

The weights for each phone were computed by measuring how well each feature identifies the phone from possible competitor phones. The next few paragraphs describe how the quality of a feature for identifying a phone from its competitors is computed.

The range of a feature, as obse.ved over all phones was first divided into N bins. In this study, N was empirically chosen to be 10 after examining the number of training tokens per phone. With too few bins, the quantized counts would show little variation. With too many bins, most of the bins would contain either 0 or 1 samples, and the counts would again be meaningless. The percentage of samples which fall into each bin was then computed for each phone. Figure 4.14 illustrates the binning of the F_1-Normalized-Position feature for /i/ and /ɔ/.

For phone pairs with very dissimilar distributions, as in Figure 4.14, many of the bins contain samples primarily from one type of phone. To determine the

dissimilarity of a pair of phones, the percentage of samples in each bin was compared. The dissimi' .rity, $d_k^{ij}(b)$, was measured as:

$$d_k^{ij}(b) = \frac{|s_{bi} - s_{bj}|}{s_{bi} + s_{bj}}$$

where s_{bi} is the percentage of phone i samples in bin b and s_{bj} is the percentage of phone j samples in bin b. From this equation, we note that when only one type of phone is present in a bin, the maximum bin dissimilarity score of 1 was assigned. If the values in a bin for the two phones being compared are equal, then the minimum bin dissimilarity score of 0 was assigned.

The quality, q_k^{ij}, of feature k for discriminating between a pair of phones, phone i and phone j, is the average of the bin dissimilarities:

$$q_k^{ij} = \sum_{b=1}^{B} \frac{d_k^{ij}(b)}{B}$$

Bins in which no phones are present were ignored. B thus is equal to the number of bins with at least one sample. The quality of a featnre for identifying a phone, q_k^i, is the average quality of the featnre for discriminating between each possible phone pair:

$$q_k^i = \frac{\sum_{j=1}^{J} q_k^{ij}}{J}$$

where J is the number of phones against which phone i may be compared.

The quality of a feature is used to weight the input of each feature geometrically:

$$Score(p_i) = \frac{\sum_k q_k^i \log S(p_i|f_k)}{\sum_k q_k^i}$$

This weighting was used because we wanted to emphasize the features which are good in identifying a phone and minimize the effect of measurement noise that nonrobust features add. That is, if only a few features are useful in the identification of a phone, input from the features which are not meaningful should be minimized.

The score for each word was computed as a function of the score of each of the component phones. Each phone was weighted by its duration in order to normalize for a varying number of phones in each path in the word lattice. For example, /θɔri/ and /faʳ/ may both be word candidates over the same region of speech. Since the number of phones in the two words is different, the score for each phone cannot simply be added; the path scores must be normalized to remove the effect that one path contains twice as many phones.

There are several alternatives for computing normalized word or path scores. The phone scores could be summed and then normalized by the number of phones. This method has the advantage that long phones are not given undue weight relative to short phones. It also has a disadvantage which is illustrated by the /θɔri/-/faʳ/ example. If /θɔri/ is normalized by the four phones composing it, then the weak fricative /θ/ receives a weight of .25. However, in /faʳ/, the weak fricative /f/ receives a weight of .5. Thus the weight given a phone is dependent on the number of other phones in a word or sentence. Furthermore, when alternate pronunciations of the same word are compared, for example, /θɔri/ and /θri/, then the /θ/ in both pronunciations should be given equal weight. By normalizing by the number of phones, the /θ/ in /θri/ receives more weight. Therefore, normalization by the number of phones has the undesirable effect that depending upon the word being verified, a phone may have a variable amount of input into the decision process. The chosen method of normalization, weighting by duration, avoids this problem, although it does have the disadvantage that short segments are given less weight. Normalizing the segments by duration, the word scores were computed as:

$$\mathrm{Score}_{word} = \sum_i \mathrm{Score}(p_i)\mathrm{D}(p_i)$$

where $\mathrm{D}(p_i)$ is the duration of phone i.

The simplifying assumption that each feature is independent was used in the scoring algorithm. This assumption was made to help insure that enough training samples were available to get good estimates since the number of samples required increases exponentially with the number of features (Duda and Hart, 1973). This technique ignores potential multivariate information, and correlations between dependent features cannot be used. For example, two sets of data may be well separated in a two-dimensional feature space, but when the data is collapsed to a one dimensional feature space, the two sets of data may overlap and cannot be separated as well. Thus results obtained using this simple technique provide a bottom line on results which can be expected if a more sophisticated verification algorithm is used.

4.3 Computation of the Ideal Word Lattice and Phone Boundaries

The task of evaluating the feature set and scoring algorithm was structured such that the performance of the verifier could be evaluated independently from the performance of the earlier components. An "ideal" word lattice, free from segmentation errors, served as input to the verifier. By using error-free input, methods for handling earlier segmentation errors, such as verification or definition of each boundary were unnecessary. Instead, the study focused on the selection of features for discrimination between similar phones and on the utility of a phone representation for evaluating word hypotheses. The results of this study serve to indicate the viability of using acoustic-phonetic features for verification.

To compute phone scores, as outlined in the previous section, the phonetic transcription of each word hypothesis and the location of phone boundaries must be known. A word lattice contains information about the word endpoints and segment boundaries. A simple procedure was used to locate the phone boundaries from the information in the word lattice. In this section, the procedures for computing the ideal word lattice and locating phone boundaries are outlined.

4.3.1 Computation of the Ideal Word Lattice

The ideal word lattice was derived by mapping the hand labeled phonetic transcription into a broad phonetic segmentation and then hypothesizing words by matching the words in the lexicon to sections of the broad segmentation. Rules were used in the mapping procedure to adjust boundaries and account for transcription labels which did not map directly into a broad phonetic class. To simulate the segmentation which would be produced by the broad phonetic classifier within a word, adjacent phones belonging to the same broad class were represented by a single broad phonetic label. For example, the /s/ in "zero" maps to "strong-fricative" and the /ɪ/, /r/, and /oʷ/ all map to "vowel." Thus in the broad phonetic representation of "zero," /ɪ/, /r/, and /oʷ/ are represented by one vowel segment and "zero" is represented as "strong-fricative vowel." Acoustic gemination was not accounted for in this mapping since adjacent phones which belong to the same broad class but occur in successive words are represented by two separate segments.

The mapping procedure modified the endpoints of the derived broad phonetic transcription to be different than the endpoints in the phonetic segmentation when sounds did not directly map into a broad phonetic class. This occurred when noise, glottalisation, voicebar, aspiration, or epenthetic silence were encountered. In these cases, the segment was arbitrarily divided evenly between the adjacent labels. This procedure produced an idealized segmentation for matching with the words in the lexicon.

In continuous speech, the location of word endpoints are unknown a priori. Words and their corresponding endpoints were hypothesised by matching each word in the lexicon against sections of the broad phonetic segmentation. All words matching a section of the broad phonetic transcription were collected to produce an "ideal" word lattice. Each of the words in the lattice contained the phonetic transcription of the word and the sequence of broad phonetic labels with associated endpoints.

The depth of the ideal word lattice, that is, the number of words in the lattice divided by the number of digits in the digit string, is statistically characterized in Figure 4.15. The average depth, which was computed before path and allophonic

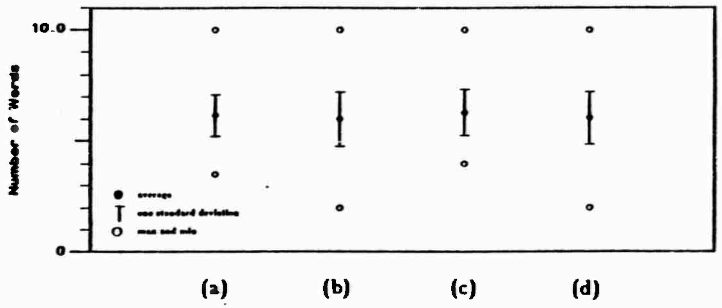


Figure 4.15: Number of Word Candidates in the Ideal Word Lattice per Word in Digit String (a) training utterances by training speakers (b) new utterances by training speakers (c) training utterances by new speakers (d) new utterances by new speakers

constraints were applied, is about 6 words/digit. This depth corresponds to a sequential constraint pruning threshold of approximately -0.5. Thus, there is much room for improvement in the performance of the broad phonetic classifier and lexical component before it approximates the ideal case.

4.3.2 Computation of Phone Boundaries

From the information contained in each hypothesized word, the verification component determined any phone boundaries not specified by the broad phonetic segmentation. It is necessary to find boundaries within a phonetic segment when multiple phones have been mapped to one segment. When the phonetic transcription of a hypothesized word contained an intervocalic /r/, an r-detector was used to locate the boundaries of /r/. The intervocalic /r/ detector first located the time, t_{min}, at which the median smoothed /r/-spectral-concentration parameter is a minimum. Anchoring from this point, the /r/-spectral-concentration parameter was searched outward in both directions, as in broad phonetic classification, until the parameter rose to 30% of the difference between the value at t_{min} and the minimum of the local maxima on each side. The front/back-spectral-concentration parameter was also used to locate the /r/ boundary when in a front vowel context. This is because the /r/-spectral-concentration parameter tends to remain low for a longer time due to the higher frequency location of F_2. The front/back-spectral-concentration parameter was searched from t_{min} until it rose to 30% of the difference between its value at t_{min} and the minimum of the local maxima on each side. The minimum distance from t_{min} to the computed edge for the /r/- and front/back-spectral-concentration parameters were defined as the /r/ boundaries. When more than one phone mapped to a broad phonetic segment and an intervocalic /r/ was not present, a default weighting, which assigns /w/ and /r/ half the duration of the other phones, was used to divide the time among the phones mapping to the segment.

4.4 Evaluation

The performance of the verifier was evaluated in terms of the word error rate and the rank of each phone in the correct path. Because the task is continuous speech, the words were evaluated subject to path constraints. That is, the "best" words were the string of words which formed the best scoring path through the word lattice. Separate evaluation of each word relative to a hand labeled orthographic transcription was not used because the meaning of comparing competitor words which have different endpoints is unclear.

The word error rate was computed by observing how often the words comprising the best word string did not match the words comprising the correct word string. With this method, a word could be the best scoring word over a region but may

not be in the best path. Insertions, deletions, substitutions and matches were computed using a 50% overlap criteria similar to that used to evaluate the broad phonetic classifier. Thus, if at least 50% of each of two words in the best word string is covered by one correct word, then an insertion is said to have occurred. Pauses between words were not included in the computation of word errors.

The best-scoring path through the word network was found using a depth-first search (Winston, 1984) without any constraints on the number of words in the digit string. The use of an exhaustive search algorithm insures that the system finds the best answer from the information it is given. Thus the effectiveness of the set of detailed features combined with the scoring algorithm was evaluated independently of any heuristics which could be used to reduce the search time.

4.4.1 Discrimination vs Identification

The discrimination and identification scoring methods were compared on a subset of the training data. The word error rate using the identification method was 3.6% and the word error rate using the discrimination method was 2.0%. As discussed earlier, the discrimination scoring method was expected to give better results because the word candidates had been reduced to a set where fine differences between competitor phones existed; thus a discrimination paradigm was more appropriate. Since the results bear out this expectation, the rest of the evaluations were performed using the discrimination scoring method.

4.4.2 Word Errors

Table 4.1 shows the word error rates for different testing conditions. Each insertion, deletion, or substitution was counted as an error. The error rate of 1.5% on the training utterances by training speakers illustrates the power of using a few carefully selected acoustic features combined with statistical measures to estimate the goodness of a phone. The error rate for training speakers on new utterances is

Table 4.1: Word Error Rates

Utterances	Speakers	# of Speakers	# of Digits	Word Error Rate
training	training	6	1365	1.5%
new	training	4	1126	5.0%
training	new	3	364	4.9%
new	new	4	893	5.3%

approximately the same as the error rate for new speakers on new utterances; this indicates that an acoustic-phonetic approach is potentially speaker-independent. The error rate for training utterances and new utterances spoken by new speakers is approximately 5%. This shows that for new speakers, the system can handle new utterances about as well as the utterances it was trained on.

A more detailed analysis of the errors in all corpora reveals that many of the errors were due to male/female differences. Some of these errors are listed in Table 4.2. The most striking and consistent error is the confusions of "four" and "five". All 16 cases in which "five" was mistakenly labeled as "four" occurred in speech by males. Eighteen of the 19 cases in which "four" was confused as "five" occurred in speech spoken by females.

Considering the acoustic differences between "four" and "five," and the differences between male and female speech, these errors can be attributed to selecting features which are not independent of male/female differences. The first phone in both "four" and "five" is /f/. Since both /ɔ/ and the initial portion of /aʸ/ are low back vowels, the coarticulation effects on /f/ due to the following vowel are approximately the same. Hence, the /f/ is similar in both words, and the main difference between the two words lies in the vocalic portion. In the vocalic portion, F_2 rises in both "four" and "five": in "four" it rises for the production of /r/, and in "five"

Table 4.2: Sample Male and Female Word Errors

Digit	Recognized as	# Males	# Females
four	five	1	18
five	four	16	0
three	four	0	7
two	sero	0	9
sero	seven	3	13

it rises during the latter portion of the /aʸ/. One of the primary differences is the higher initial location of F_2 in /aʸ/ than in /ɔ/. However, for the same vowel, the location of F_2 varies among speakers. It is generally higher in frequency for female speech than male speech (Peterson and Barney, 1952), since females have a shorter vocal tract length. The Spectral-Offset-Location was designed to be sensitive to differences in the initial location of F_2 in /aʸ/ and /ɔ/. Considering this sensitivity, the errors of labeling the /ɔ/ in "four" as an /aʸ/ in female speech, and the /aʸ/ in "five" as an /ɔ/ in male speech, are reasonable. A better, speaker-independent feature may be to compare the Spectral-Offset-Location relative to the location of F_3, since a larger dip in the spectrum is observed in /ɔ/ than in /aʸ/.

To obtain an idea of the robustness of the verification scores, the score of the correct word relative to the score of the other word candidates was examined. In particular, the score of the top candidate was compared to the score of the second best candidate when the top candidate was correct. When the top candidate was incorrect, its score was compared to the score of the correct word. Figure 4.16 illustrates this for each of the test sets. Note that the difference in word scores is generally small when an incorrect word is the best scoring word, and that the difference has a large range when the correct word is the best scoring word. In a recognition system, this information could be used to reject the utterance when

Table 4.3: Phone Rank in Correct Words

Speakers	Utterances	Position			
		0	1	2	3
Training	Training	.93	.99	1.00	1.00
Training	New	.86	.98	1.00	1.00
New	Training	.90	.99	.99	1.00
New	New	.86	.98	.99	1.00

the difference in word scores is small, and the speaker could be asked to repeat the utterance. Alternatively, this information could be used to identify words which do not score much better than their competitors, and finer discriminations could be performed on these words.

4.4.3 Phone Errors

The rank of each phone in the correct word was also used to evaluate the verifier. These results are shown in Table 4.3 for the test sets. Note that for new sentences by both the training speakers and the new speakers, the correct phone is in the top position at least 86% of the time and within the top two candidates at least 98% of the time. This similarity in rank indicates the potential speaker-independence of using acoustic features for verification. As expected, the largest percentage of phones were in the top position when the system was tested on the test set composed of training utterances by training speakers. However, the top two ranking candidates include the correct phone at least 98% of the time on all corpora. These results indicate the viability of performing verification at the phone level using acoustic-phonetic features.

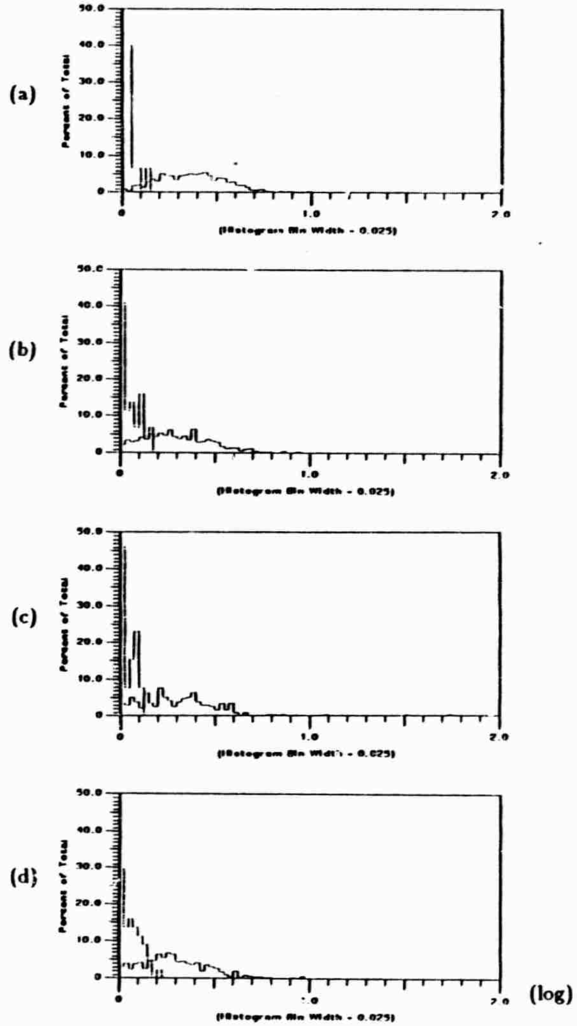


Figure 4.16: Difference in Word Scores between Correct Word and Next Best Word (solid line) and Best Word and Correct Word (dashed line) for: (a) training utterances by training speakers (b) new utterances by training speakers (c) training utterances by new speakers and (d) new utterances by new speakers

4.5 Chapter Summary

The main points of this chapter are:

- Words were scored based upon the value of phone features, demonstrating a potentially extendable scoring method.
- Use of a phone representation allows features to be defined over regions which are minimally affected by coarticulation and also allows a wider variety of characterisations of speech to be exploited.
- Better performance was achieved using discrimination between phone competitors than identification based purely on featnre values.
- Verification using acoustic features is potentially speaker-independent.
- A small set of well chosen acoustic features is adeqnate for verification of phones in the digit vocabulary.

Chapter 5

Discussion

This chapter discusses the nnderlying assumptions and contribntions of this thesis research, which was focnsed on examining how low-level acoustic-phonetic speech knowledge can be used in continuous speech recognition. Several underlying assumptions were made in this research. One of the basic assumptions was that an acoustic-phonetic approach is worth exploring. It was also assumed that speech can be represented as a sequence of sounds associated with regions of the speech signal, and that a phone is a good unit for representing speech. The following sections attempt to clarify the reasoning behind these assumptions and the choice of the digit vocabulary as a case study. Additionally, the merit of each component in the acoustic-phonetic recognition model is considered. The use of a preprocessor, in particular, a preprocessor based on acoustic-phonetics, in a recognition system is discussed. The reasons behind the choice of a simple control strategy for studying how detailed acoustic features can be used in verification is addressed. Computational requirements of a segment-based approach are also considered. Finally, the contributions of this thesis towards a better understanding of the use of acoustic-phonetic knowledge in speech recognition is outlined, and suggestions for future work are made.

5.1 Why an Acoustic-Phonetic Approach?

An acoustic-phonetic approach is appealing both intuitively and in its potential extendability to less restrictive recognition tasks. It is intuitively appealing because it provides a framework for describing speech sounds and coarticulation and also for applying such knowledge in a recognition system. Since the knowledge used by the system is specified using human knowledge, its application is often under explicit human control.

By incorporating speech knowledge–knowledge about the problem domain–the errors which the system produces are usually reasonable. That is, one can usually understand why the errors occurred. Therefore, the errors may be corrected by extending the knowledge. For example, when computing phone scores based on how likely the feature values indicate that a phone is the hypothesized phone, many of the errors appeared to result from differences between male and female speech. By choosing features which are less dependent on these differences, or by normalizing the values based upon speaker characteristics, these differences may be more readily handled.

In contrast, other approaches do not provide a way to explicitly handle such errors. These approaches incorporate speech knowledge only implicitly in the training data or by folding the information into the recognition algorithm. Thus the system must either be retrained or the recognition algorithm must be modified to "correct" errors which occur only in a subset of phones. In template matching, researchers have the option to increase the amount of training data and the number of reference templates used. However, this has the undesirable consequence that attempts to correct errors which occur with only one type of speech sound increase the number of templates for all recognition units. In addition, researchers must explicitly collect data containing the variation or else the new templates may be based upon only a few outliers, resulting in non-robust templates.

An acoustic-phonetic approach is also appealing in its potential extendability to

larger vocabularies and more complex task domains. The extendability results from the use of a phone based representation. The number of phones in any language is limited. English uses about only 40 phones; hence, the maximum possible number of diphones, or phone pairs, is limited to about 40^2. Furthermore, the number of consonant phone sequences within a syllable is finite. Studies have been conducted on the number of unique consonant phone sequences in a subset of commonly occurring English words. These studies show that as new words are added, new phone sequences occur, as expected. More importantly, these studies show that the possible number of allowable sequences within a syllable is limited. In particular, only about 70 syllable-initial and 130 syllable-final consonant sequences exist in English (e.g., Shipman and Zue, 1982). Consequently, the maximum number of contexts in which a phone can occur is much less than the number of words in English. Since each word can be represented as a phone sequence in an acoustic-phonetic approach, and since the number of recognition units is independent of the number of words in such a representation, an acoustic-phonetic approach is potentially extendable.

A considerable investment is required to develop the knowledge needed to characterize phones in different environments for a phonetically based system. However, once this investment has been made, the vocabulary should be extendable simply by adding one or more phonetic representations for each new word. In contrast, a word-based system must train each new word, possibly in each environment in which it could occur. Thus, we believe that the investment needed to gather knowledge for the development of acoustic-phonetically based systems has large potential payoffs by providing a framework for exploiting speech knowledge and for developing less restrictive speech recognition systems.

5.2 Why the Use of Digits as a Case Study?

Selection of the digits as a case study for exploring constraints in continuous speech recognition is the result of our attempt to demonstrate the model, and at the same time to keep the problem manageable. Admittedly, the digit vocabulary is limited in many ways. For example, syntax and semantics are not applicable to random-length digit strings. In addition, stress is not of primary importance because the amount of stress given each word in a digit string is approximately the same: there are no function words in a digit string, each word is of equal importance, and most words are monosyllabic. Furthermore, the digits do not demonstrate the phonetic richness found in American English.

However, the digits form a suitable vocabulary for studying how each component in our model could be implemented to handle the variations which occur in speech. Although the digit vocabulary does not include all the phones in English, it contains examples of allophonic variations and many low-level phonological effects, such as acoustic gemination and flapping. The phonological effect of acoustic gemination, that is when two similar phonemes merge into one and appear acoustically as one segment, is observed in both consonants and vowels in digit strings. For example, the final /s/ is "six" geminates with the initial /s/ in "seven," such that the two /s/'s appear acoustically as one /s/. The final /s/ in "six" also geminates with the /z/ in "zero." The low-level phonological rule for flapping of /t/'s in an intervocalic position is also illustrated in the digit vocabulary. In the digit strings, the /t/ in "eight" is in intervocalic position when followed by another "eight," and this intervocalic /t/ can be flapped.

Some allophonic variations are observed in digit strings. For example, three realizations of /t/—released, unreleased, and flapped—occur. The released /t/ is usually observed word initially, as in the word "two," or word finally when the word is the last word in a phrase, as in the word "eight" at the end of a sentence. A /t/ in word final but not phrase final position, as in the word "eight" in the middle of a digit string, is frequently unreleased. And a /t/ may be flapped when it occurs in intervocalic position, as in the string "eight eight." At least two realizations of

/v/ are also observed. A strongly fricated /v/ often occurs before a fricative, and a strongly voiced /v/, in which the frication rides on the voicing, occurs primarily in intervocalic positions. In addition, the nasal /n/ occurs in both intervocalic and non-intervocalic position. Thus both robust intervocalic nasals and generally weaker non-intervocalic nasals are illustrated in the digit vocabulary.

The digit vocabnlary also illustrates many coarticulation effects, both within words and across word bonndaries. The second formant in "two" is raised initially due to the preceding /t/, which is a dental consonant, and it may remain raised if followed by another dental, such as /s/, or it may drop very low in freqnency if followed by /w/. Similarly, the formant values of the word final /o^w/ in "zero" are strongly affected by the following phonetic environment. Thns the digit vocabnlary illustrates a wide variety of phonological variations and coarticulation effects which must be handled by a recognition system.

The digit vocabnlary is suitable for demonstrating the utility of the acoustic-phonetic model components. A vocabulary which illustrates the full power of sequential constraints in word candidate reduction leaves little work for the verifier. Consequently, the role of the verifier cannot be studied. For example, the sequential constraints in a small polysyllabic vocabulary may be so strong that most of the word boundaries and most of the words can be specified without ambiguity. Since most of the words are specified, the verifier is not needed except in a few cases. In contrast, the digit vocabulary is primarily monosyllabic and most of the pronunciations of each word contain very few broad phonetic segments. Since some sequential constraints apply, the utility of sequential constraints in recognition can be studied using the digit vocabulary. And since the digit vocabulary does not illustrate the full power of sequential constraints, the role of the verifier can be studied.

In summary, many examples of phonological variation occur in the digit vocabnlary, even though it is limited in many ways. Furthermore, the digits form a manageable task for demonstrating how acoustic-phonetic knowledge can be applied to speech.

5.3 Why a "Preprocessor" Based on Acoustic-Phonetics?

In our recognition model, the broad phonetic classifier and word hypothesizer can be view as a "preprocessor." The purpose of the preprocessor is to rule out unlikely word candidates based upon information that is easy to compute. A preprocessor has been used in other recognition systems as well. For example, an isolated word recognizer developed by Pan et al. (1985) used a preprocessor based on vector quantization (VQ) (see for example, Buzo et al., 1980) to screen word candidates for recognition using dynamic time warping. The primary purpose in their use of a preprocessor was to reduce computational costs while maintaining the performance rate of current DTW systems.

However, an acoustic-phonetic preprocessor has many advantages in addition to possible reduction of computational costs. By using an acoustic-phonetic preprocessor to apply speech constraints to remove poor word candidates, more directed detailed acoustic analysis can be performed. Additionally, an acoustic-phonetic preprocessor segments the speech so the benefits associated with a phonetically-based segment representation can be exploited. These advantages are discussed in more detail in Section 5.4 and Section 5.5.

In our model, an acoustic-phonetic preprocessor screens word candidates based upon broad phonetic information, and the word candidates are evaluated based upon robust information. In contrast, a non-acoustic-phonetically based preprocessor attempts to screen word candidates primarily on detailed spectral information. The purpose of the preprocessor is to rule out unlikely word candidates; attention to fine phonetic differences at an early point in processing is not only unnecessary but is also not as robust. Fine phonetic differences are not as robust because fine

phonetic differences are more sensitive to allophonic variation than differences between a broad phonetic class. This contrast is important because the preprocessor removes word candidates, and the recognizer may not provide for recovery of word candidates which should not have been removed. Thus the decision threshold for removal of a word candidate should be lenient to avoid irrecoverable errors. However, if the information given the preprocessor is not very robust, then the preprocessor can not be efficient in word candidate reduction.

An acoustic-phonetic preprocessor is applicable to continuous speech, as demonstrated in this thesis. In contrast, other preprocessors, such as the VQ-based preprocessor are word based. Therefore, extension of these preprocessors to continuous speech is uncertain because unknown word endpoints must be dealt with.

An acoustic-phonetic preprocessor allows different types of speech information to be incorporated in the identification of broad phonetic classes. Pan et al. found that a VQ preprocessor performed much better when temporal and energy information were also used. However, the information was incorporated into the existing time-frame structure, which is not conducive to using such information.

Most recognition models which do not use an acoustic-phonetic preprocessor recognize an utterance by matching a set of templates to the input signal. Three difficulties that are associated with this recognition method, but are minimized by using an acoustic-phonetic preprocessor are: (1) speaker independence is not easily incorporated, (2) context cannot be explicitly used, and (3) the speech signal is quantized to the template values used in a template representation.

Models which perform spectral matches are inherently speaker-dependent because the spectral representation has not been abstracted to capture the speaker-independent information. DTW and VQ attempt to handle speaker-independence with use of multiple templates. In a network model, multiple paths may be needed to represent different types of speakers. Thus each approach attempts to achieve speaker-independence by capturing variations in sounds, rather than abstracting the speaker-independent features of sounds, as is possible in an acoustic-phonetic approach. By capturing variations in sounds, measurements made in these approaches are inherently noisy. This is because in addition to information relevant to the speech sounds being identified, information irrelevant to the speech sounds being identified is incorporated into the measurement.

By not using a preprocessor to find a speech motivated recognition unit, these models must use a regularly sampled representation of the input signal. However, a uniformly sampled representation has the undesirable quality that context is difficult, if not impossible, to specify. This is because context is a speech based concept and not a time-frame based concept. In contrast, an acoustic-phonetic preprocessor allows context to be explicitly specified.

Representation of a rising F_2 by a network model with a fixed number of states per phone illustrates the quantization problem. A network model, if it has a sufficient number of states (e.g. Harpy), would try to capture *rise* in F_2 through a sequence of states. However, the quality of match as the speech signal passes from one state to the next in the sampled representation varies. This variation is not due to variations in the quality of the rising F_2; it is due to quantization in the match. Thus measurements made in network models without a preprocessor do not accurately reflect the events occurring in the speech signal. In contrast, the segments defined by an acoustic preprocessor allow the movement of a formant to be explicitly captured.

In summary, an acoustic-phonetic preprocessor is a valuable part of a recognition system. The preprocessor uses robust information, is applicable to continuous speech tasks, and allows different types of speech knowledge to be easily incorporated into the recognition process. It also provides a basis for performing speaker-independent recognition, allows context to be specified, and allows a more accurate representation of events in the speech signal.

5.4 Why Segments?

An assumption made in this work, and an integral part of the approach is that a sequence of labels may be attached to the speech signal. Furthermore, it was assumed that speech is produced as a sequence of sounds of varying duration which can be represented as a sequence of labels. The sequence of labels, or recognition units, correspond with important phonetic events. As a result, the phonetic recognition units will be irregularly spaced. A phonetic unit can be associated to a (perhaps fuzzy) time in the speech signal or to a (perhaps fuzzy) region of the signal.

This thesis uses phonetic units which were associated with regions of the speech signal, and each region is referred to as a phone segment. This section addresses two issues. First, a segment representation is argued to be superior to a time-frame representation. Second, the advantages of a phonetic segment representation over other segmental representations, such as the diphone and demisyllable, are described.

A segment representation has many advantages over a time-frame representation. For example, a segment which spans a region of speech can be characterized over time. This is an important attribute because many strong cues to speech sounds are distributed across time. For example, systems based on segments or a sequence of phonetic units can explicitly characterize formant motion during the first 30 msec of a vowel to capture transition information. In a frame-by-frame analysis, this information is included, along with other information, only implicitly in the training data.

The use of segments rather than individual spectral analysis frames allows a wider variety of acoustic-phonetic constraints to be exploited. That is, in addition to the ability to characterize transition and relatively stable regions of the speech signal, characterizations over the entire region, such as the maximum, minimum, or average value of a parameter, are available. With the segment formulation, onset rate can be used to influence the decision on the identity of the whole segment. In contrast, features such as onset rate do not make sense in a frame formulation and would only influence the score in one frame of a spectral distance metric, if at all. Thus a segment framework allows important information to be taken into account explicitly.

As a consequence of the variety of characterizations available with segment-based representations, a system which uses a segment representation can avoid many of the errors produced by systems which simply try to match the spectrum. This is because there is much more information in the speech signal than spectral shape. For example, one striking difference between a strong alveolar fricative and weak dental fricative, given the same context, is the strength of the fricative. Distance metrics such as Itakura's (1975) do not use energy information. Instead, such information must be explicitly incorporated if it is to be used in recognizers based on spectral distance formulations. Researchers are now recognising the importance of using such information and are devising methods for incorporating such knowledge into existing algorithms, such as vector quantization (Pan et al., 1985; and Bush and Kopec, 1985) and Hidden Markov Modeling (Schwartz et al., 1985). The use of features in an acoustic-phonetic approach provides a unified method for using this knowledge. In an acoustic-phonetic approach, features may be selected based upon human knowledge of what is important, supplemented by statistics to verify that the knowledge has been adequately captured by computer. Many speech motivated segment representations, such as the phone (e.g. Woods et al., 1976), diphone (e.g. Scagliola and Marmi, 1982), syllable (Fujimura, 1975; Mermelstein, 1975), and demisyllable (Rosenburg et al., 1983), have been proposed. We believe that a phonetic representation is better than either the diphone or demisyllable representation. It is more flexible because it can be transformed into a diphone or demisyllable representation. Therefore, all the information which is available from a diphone or demisyllable representation is also available from a phone representation.

Furthermore, many characteristics which are easily computed from a phonetic segment representation are more difficult to extract in a diphone or demisyllable based representation. For example, measurement of duration, such as the duration of a fricative, is straight-forward in a phone representation. In contrast, in a diphone representation, each phone has been split into two parts to form diphones; as a result, information about phone boundaries and phone durations are not easily obtainable.

Acoustic-phonetic knowledge, such as contextual information, can be easily and explicitly expressed using a phone representation. A phone representation is amenable to using information during the relatively stable central portion of the phone and also to using transitional information. This is possible because the phone unit defines regions of the signal which should be stable and also points of transition (the edges of the region). Since acoustic features may be defined over any region of a phone, features characterizing transition regions and features characterizing stable regions can be defined. For example, the feature characterising the average value of F_1 was computed over the middle 50% of a phone. Since most of the contextual information is contained in the transitions at the beginning and end of a phone, this estimate of F_1 is minimally influenced by context.

By making judgments about sounds based on characteristics over a region of the signal which is generally stationary, such as within a phone, the effect of local variations in the signal can be reduced by techniques such as averaging. Furthermore, estimates of values characterizing the region should be more accurate than the ensemble of single estimates for each point, since the value characterizing a segment is based on examining the data in the region as a whole. Thus the judgments made over a region should be more reliable.

In summary, the use of segments allows a wide variety of acoustic-phonetic constraints to be exploited and a wide variety of characterizations of the speech signal. The flexibility in specifying the importance of information in the signal, available when the signal is represented by segments, eliminates many errors encountered in spectral matching systems. A phonetic segment representation is superior to other representations because it allows characterization of information available in a diphone or demisyllable representation; furthermore, it allows characterization of other information derived from phonetic units.

In this section we have argued the advantages of a segment representation. However, we should note that we are imposing a segment representation on the speech signal and that a segment representation is a convenience which we use to describe the speech signal.

5.5 Why Use Sequential Constraints in Lexical Access?

Lexical access is the point in the recognition process where information about the speech signal is combined with knowledge about the vocabulary to propose word candidates. Sequential constraints provide a mechanism for removing unlikely word candidates from consideration before the fine discrimination necessary for identification of a phone is performed. The remaining word candidates are relatively similar since each candidate is composed of a string of phones which match the string of broad classes well and therefore match the initial features characterising a broad class well. As a result, features used initially to identify broad classes are not needed in verification, allowing the verification component to perform more directed analyses.

If sequential constraint application is skipped so that all words are hypothesized at each possible position, then the verifier is burdened with additional phones and words to score. Additional features may be needed since fine as well as gross differences between sounds must be measured, increasing the amount of computation. The contrast between two similar competitors also is reduced; noise may be added to the measurements because many other unlikely competitors, such as c in Figure 5.1, are also compared to "real" candidates, *a* and *b*, thus adding an "offset" to the scores.

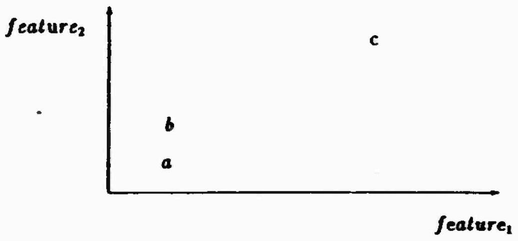


Figure 5.1: Real competitors *a* and *b* and outlier c

The use of sequential constraints in lexical access thus allows the verifier to focus on the differences between two similar competitors, rather than measuring that *a* and *b* are similar relative to c, and also measuring how much better *a* is to *b*. As an analogy, if one is trying to measure the peak-to-peak amplitude of an ac signal which is offset by a large dc bias, one would measure on a scale covering only the ac region to get an accurate measurement; one would not include the dc component in the measurement.

In the application of sequential constraints, a risk is associated with ruling ont correct word candidates. However, recognition systems based on network models also implicitly use constraints in the search strategy. In continuous speech, the possible paths in a network model based on frame-by-frame analysis can be so large that searching the entire network is intractable. These search strategies generally employ some type of heuristics for pruning, such as the beam search used by the Harpy system. To make a contrast, pruning is applied while searching in a network model, but pruning is applied before creation of a network to be searched in an acoustic phonetic model. The use of speech constraints, such as sequential constraints, is advantageous over heuristic search techniques because the risks associated with using speech based constraints can be quantified. However, unless the speech constraints are applied simultaneously, the use of speech constraints cannot be guaranteed to rule out only words which can never achieve a better score.

A network is used to represent the phones in the verification component. The phones in a word are associated with the transitions, and the cost of making a transition is a score which reflects the "goodness" of a hypothesized phone, based upon feature information. The word scores from lexical access may be used in this network representation in several ways. Based directly on the results for using sequential constraints, a pruning threshold can be set to limit the number of word candidates considered. A network can then be constructed from the remaining word candidates. Alternatively, if a heuristic search strategy is used where nodes are expanded as needed, the words candidates could be stored in an ordered queue based on their lexical access score. Nodes could then be expanded using words from the ordered queue. The disadvantage of this method is that the set of phone competitors can change as nodes are expanded. Consequently a phone cannot be scored using discrimination between a phone and its competitors; instead, scores must be assigned based purely on identification information.

In summary, the use of sequential constraints allows the verification component to perform a more detailed and directed evaluation of the phone and word candidates. The risks associated with using speech constraints can be quantified, something which is not as easily achieved using heuristic search algorithms. When constraints are applied to reduce the word candidates at once, competitor phones are known and discrimination between phones can be performed.

5.6 Why a Simple Control Strategy?

The verification component used a simple control strategy: find the best sequence of digits by maximizing the score of each phone, where the phone scores are a function of the probability of a hypothesized phone occurring, given the observed feature values. This simple strategy was found to be sufficient for verification of phones in the digit vocabulary.

In a complicated recognition task, a control strategy may be needed which integrates cues in a logical manner and hypothesizes new phones if conflicting cues indicate that more than one phone is present in a region which originally was thought to represent only one phone. An example of a situation in which a new phone should be hypothesized is when clear labial formant transitions are apparent in the left context of a closure, followed by a compact double burst located slightly above F_2 (strong indication of a velar release). Instead of a single stop, two consonants should be hypothesized as occurring over the region: a labial consonant, such as /p$^{\urcorner}$/, /b$^{\urcorner}$/, or /v/, followed by a velar stop.

Rather than developing a complex control strategy as outlined above, the simple control strategy was chosen because our understanding of the use and integration of acoustic cues is still very primitive. Furthermore, we are still trying to devise algorithms for effectively capturing the acoustic cues needed by a more complex control strategy.

Thus the emphasis in implementing the verifier was placed on assigning identification scores. Although the earlier example (where two different consonants were hypothesized) is not applicable to the digits, potential conflicts can occur in the digit vocabulary. For example, when acoustic gemination occurs, only one broad phonetic segment is found, but two phones should be hypothesized. Potential conflicts which require additional segments to be hypothesized were avoided by initially hypothesizing all viable candidates. Durational and sequential constraints were used to determine whether a candidate was viable. An identification score for each hypothesized phone was then computed based on the value of well-motivated features and statistics.

The feature values were weighted based upon how well each feature identified the phone in question. Thus a feature which is not relevant to the identification of a phone will be given little weight and minimally affect the score. When a feature indicated the possibility of a hypothesized phone to be contrary to what the other features indicated, this was reflected in the computed score. For example, all features may indicate the possibility of an /s/ to be high, except for one which strongly indicates that an /s/ is very unlikely. The resulting score of the hypothesized /s/ is reduced by an amount which depends on how heavily the conflicting feature is weighed.

The assignment of feature weights assumed that the value of each feature varies over a range, rather than being a conditional value. For example, the value of the feature F_1-Normalized-Position ranged from -1.0 to 1.0. A conditional feature such as whether or not a double burst is observed does not satisfy this assumption. A double burst in the region of F_2 is a strong indicator of a velar release (Cole, et al., 1980). Thus one would like to weight this information conditionally such that when a double burst is observed, the feature capturing this phenomena wonld be given a very heavy weight. When a double burst is not observed, the featnre would be given no weight, and the stop would be identified using other features, such as burst location. The current approach can be extended to handle this type of phenomena.

A sophisticated control strategy would use information about coarticulation in making decisions. The strategy used instead was to select features which are least affected by coarticulation, such as making measurements in the middle 50% of a phone. This strategy was shown to be sufficient for handling coarticulation in the digits.

5.7 Computational Considerations

Recognition systems are currently limited in part by the amount of computation required, and future recognition systems will have even larger computational requirements if current technologies are simply extended. We believe that an acoustic-phonetic approach has the potential to be used in less restricted tasks because of the types of constraints which are available by using this approach.

Computational requirements can be reduced in an acoustic-phonetic approach by applying speech knowledge to rule out unlikely word candidates. It was shown in this thesis how a particular speech constraint, sequential constraints, can be effectively applied. Even when given input from a front end which produced a non-ideal segmentation, sequential constraints were found to provide a reduction in the number of word hypotheses when the input error characteristics of the front end are known and used. Furthermore, this candidate reduction was achieved with new (broad phonetic) pronunciations of words.

With a phone segment representation, the maximum number of recognition units is limited, and more importantly, the maximum nnmber is independent of the number of words in the vocabulary. Thus the use of a phone representation in an acoustic-phonetic approach gives the approach the potential to be computationally tractable in more unrestricted recognition tasks.

The current implementation used a word spotting approach which is not computationally efficient. This is because the focus of the study was quantification of the constraint provided by knowledge of the lexicon, without interaction from heuristic search techniques, in contrast to the development of a recognition system. However, in implementing a verifier as part of a recognition system, the required computation could be restricted by use of sequential constraints in conjunction with efficient search techniques which expand nodes as needed.

In an acoustic-phonetic approach to continuous speech recognition, the phone endpoints are located before the computation of recognition scores. Thus recognition scores are computed for only one set of endpoints. This approach can be contrasted with template-based approaches which try to find the best set of endpoints by computing a recognition score for each possible endpoint pair. As the vocabulary size increases, the number of possible endpoints which must be considered in a template matching approach increases. Therefore, since recognition scores need not be computed multiple times for different sets of endpoints, the computation required by an acoustic-phonetic approach is more tractable than a template matching approach. We believe that through the use of speech constraints and a segment representation, an acoustic-phonetic approach is a computationally tractable approach for the development of less restrictive speech recognition systems.

Summary

The previous sections have addressed the following issues:

- An acoustic-phonetic approach is appealing intuitively and in its extendability to less restrictive tasks.
- Digits were chosen for the case study because many phonological variations in speech occur in digit strings. Thus the digit task allowed the utility of the model to be demonstrated, yet the constraints on the vocabulary kept the problem manageable.
- An acoustic-phonetic preprocessor is a valuable part of a recognition system because it allows a phone-based segment representation, explicit specification of context, and a more accurate representation of events in the speech signal.
- Segments allow a wider variety of characterizations of the speech signal, eliminating many of the errors encountered in spectral matching systems.
- Sequential constraints reduce the burden on the verifier and allow verification to be more directed and detailed. In addition, the risks associated with application of sequential constraints can be quantified so that they are known.

- A simple control strategy using acoustic-phonetic features based on speech knowledge is sufficient for the digit vocabulary.
- Because of the types of constraints available in an acoustic-phonetic approach, we believe that this approach has the potential to be used in less restricted recognition tasks.

5.8 Contributions of the Thesis

This thesis illustrates that an acoustic-phonetic approach is a viable alternative for building continuous speech recognition systems. This was demonstrated by extending the acoustic-phonetic recognition model for isolated words proposed by Shipman and Zue to continuous speech and by implementing the components in the model for the digit vocabulary case study. Implementation of the components demonstrated how speech knowledge could be used in each component. More importantly, implementation of the components allowed study of the issues involved with handling speech variability when applying speech constraints, a better understanding of how speech constraints can be applied, and a better understanding of how speech constraints and an acoustic-phonetic approach can be used to reduce some of the primary difficulties in developing a speaker-independent, continuous speech recognition system. Although this research used the digit vocabulary as a case study, the observations and issues addressed should be of value in the design of continuous speech recognition systems using other vocabularies.

5.8.1 Extending the Theoretical Model

In this thesis, it was demonstrated that it is feasible to extend the Shipman and Zue model to continuous speech for a limited vocabulary such as the digits. Shipman and Zue's model proposed that sequential constraints applied to *broad class* information from the speech signal provides a significant reduction in the number of word hypotheses. To extend the model to continuous speech, broad class sequential constraints are used to hypothesize words and also to hypothesize corresponding word boundaries. Performing word hypothesis from a broad phonetic segmentation is in contrast to earlier continuous speech recognition systems, such as HWIM and Hearsay II, which hypothesized words from a phonetic input string. The feasibility study demonstrated that in the digit vocabulary with multiple pronunciations of a word allowed, 66% of the word boundaries could be identified given an error-free input. Furthermore, using path constraints and sequential constraints, an average of 2.8 digits were proposed for every digit in the string. These results show that a significant reduction in word candidates can be achieved using sequential constraints, and, as a result, a recognition model based on sequential constraints can be used for continuous speech.

5.8.2 Contributions from Component Implementation

We explored an acoustic-phonetic model of continuous speech recognition by implementing the model components. By taking the step from a proposed model to implementation of the components of the model, issues important to the application of speech constraints in a recognition system were studied. A method which used front end characteristics for applying sequential constraints to non-ideal data was developed and shown to provide effective candidate reduction. In addition, implementation of the verifier demonstrated the power of an acoustic-phonetic approach which allows a few well-motivated acoustic features to be selected for identification of phones.

Broad Phonetic Classifier

Implementation of the broad phonetic classifier demonstrated several concepts. It illustrated that a set of acoustic features describing robust characteristics of broad classes of speech could be identified. Furthermore, these features could be combined

using a set of production rules to produce a broad phonetic segmentation. Implementation also demonstrated an alternative to earlier segmentation algorithms. Rather than assigning labels to each frame and then grouping the labels to form segments, robust regions, similar to islands of reliability, were identified and then extended outward.

Word Hypothesizer

Implementation of the word hypothesizer illustrated how sequential constraints can be effectively applied to speech. In conjunction, two issues were addressed: (1) Is a segment lattice or a segmentation string a better representation for input to the word hypothesizer? and (2) Can knowledge about front end characteristics and segmentation of speech be combined to produce a score indicating the viability of each word hypothesis?

The variations which occur in real speech dictate that flexibility is required to apply sequential constraints. The segment lattice and segmentation string represent two approaches to handling front end characterizations of speech variations. The segment lattice attempts to handle variations by allowing multiple labels. However, the ambiguity of the lattice makes computation of a meaningful characterization of the broad phonetic classifier questionable. Thus when sequential constraints were applied to a segment lattice, knowledge about the characterization of the broad phonetic classifier was not used. It was found that this method of lexical access did not provide enough flexibility to handle new broad phonetic representations of words, even when a lexicon of alternative pronunciations was used.

In contrast, the insertion, deletion, and substitution characteristics of the broad phonetic classifier can be defined from a segmentation string. A scoring algorithm which computed how well the phonetic pronunciation of a word matched a portion of the segmentation string was developed. The algorithm generally penalized new but similar pronunciations of a word only slightly. The distribution of correct word scores were observed to cluster near a probability of 1. In contrast, the distribution of the scores of all word hypotheses was very broad. It was shown that these results could be used to set a threshold as an effective way of removing poor word candidates from consideration by the verifier. At the same time, words which had a slightly different broad phonetic representation than the canonical representation derived from the phonetic transcription generally were not ruled out, allowing for new but similar pronunciations of words.

It was observed that path constraints and broad allophonic constraints are not as effective when the number of word hypotheses is large. Thus, in addition to permitting the verifier to be more directed, the use of sequential constraints was found to permit these constraints to be more effective.

Verifier

Implementation of the verification component illustrated that phones are a viable and useful representation. It further illustrated how, in the digit vocabulary, a few (nine) well motivated acoustic features combined with statistical characterizations can be effective in identifying phones. This success was due to the ability to selectively characterize portions of phone segments and also to the ability to variably weight acoustic events which cannot easily be given much importance in a frame-by-frame algorithm. These results thus illustrate the utility of an acoustic-phonetically based decision process. Additionally, it was found that detailed discrimination between competitors provided better identification scores than deriving a score based solely on the characteristics of the phone itself.

5.8.3 Advantages of an Acoustic-Phonetic Approach

Implementation of the model components illustrated how an acoustic-phonetic approach is potentially speaker-independent. In addition, implementation illustrated how an acoustic-phonetic approach can be used to reduce some of the common difficulties in the development of large vocabulary continuous speech recognition systems, such as coarticulation and speaker-independence.

Speaker-independence

Speaker-independence was examined in each component by comparing the output when the component was tested on new speech by the training speakers and when tested on new speech by new speakers. The outpnt of the broad phonetic classifier was found to be very similar for the two sets of speakers, illustrating that broad phonetic classes can be found independent of speaker.

The lexical access component is affected by speakers in the broad phonetic representation of the speech signal. Since the output of the broad phonetic classifier is relatively speaker-independent, the distribution of the correct word scores, based upon the automatically computed broad phonetic transcription of the speech signal, should also be similar, unless the types of insertions, deletions, and substitutions which occur have changed significantly. It was found that the distribution of correct word scores is essentially the same for training speakers and new speakers on new digit strings. This demonstrates that sequential constraints, when combined with information about front end characteristics, can be used to score word candidates, independent of speaker.

Speaker-independence in the detailed acoustic analysis component was evaluated based on phone and word recognition rates. Comparison of the rank of the correct phone in a word between training speakers and new speakers on new sentences shows no significant degradation. Similarly, comparison of the word error rate between training and new speakers on new speech is essentially the same. Thus, in an acoustic-phonetic approach, one can choose features which key on important information in the speech signal, relatively independent of speaker.

In summary, all three components were found to perform similarly for training speakers and new speakers. Thus an acoustic-phonetic approach shows promise for speaker-independent recognition.

Coarticulation

Another difficulty in the development of a continuous speech recognition system is coarticulation between words. The use of phone segments as a verification unit allows the system to examine or not examine coarticulation effects. In this thesis, identification of phones by examining regions least affected by coarticulation was studied using the digit vocabulary. It was found that phones can be identified reasonably well using information during the region least affected by coarticulation. In particular, the correct phone was the top candidate at least 87% of the time, and within the top two candidates at least 98% of the time. These results show that use of a phone representation is a powerful method for reducing coarticulation effects to derive a baseline phone identification score. Further work to explicitly exploit transitional information could thus use the best ranking phone candidates as a starting point for verification of possible transitions between phones, based on the acoustic information present.

5.9 Future Work

We believe that the results of this research indicate the viability of an acoustic-phonetic approach to continuous speech recognition and that further studies should be pursued using this model. We also believe that the next step in this area of research is to use a more formalized approach to incorporate speech knowledge into each system component. In this section, we outline suggestions towards a more formalized approach and propose modifications for each of the system components based upon what was learned through implementation of the components.

5.9. Broad Phonetic Classifier

Before using the broad phonetic classifier in a recognition system, its performance should be improved. Ideally, the performance should approach a level such that the number of word candidates remaining after application of sequential constrainte is comparable to the number of candidates in the ideal word lattice (as was used to evaluate the verifier). The broad phonetic classifier can be improved and extended by using a more formalized approach and by refining the chosen set of broad phonetic labels. Speech knowledge was incorporated into the broad phonetic classifier using heuristics. More formalised methods for defining the set of broad classes used in initial labeling and for identifying segment boundaries and labels are needed. For example, broad phonetic classes could be selected based upon how the phones in a labeled set cluster in a chosen feature space. In addition, a formalized method for identification of segments should still adapt to new utterances; that is, the classification algorithm should use training data as a standard which can be adjusted to utterance characteristics.

The chosen set of broad phonetic classes should be examined in more detail and the least robust classes refined. For example, rather than trying to identify nasals in all contexts, only intervocalic nasal consonants (which are much more robust than word initial or word final nasals in the context of a voiceless consonant) could be included in the sonorant broad class; and non-intervocalic nasals could be included in the vowel broad class. The rules for finding a short voiced obstruent should also be refined. Although a dip in energy is usually observable in higher frequencies, this information was not used; this information could be used by looking for a dip in energy in the higher frequencies and/or by looking at a wider energy band.

The broad phonetic classifier currently segments and labels the speech signal into six broad phonetic classes. By simply extending the approach to larger vocabularies, the number of word candidates will generally increase. Inclusion of several more detailed, but still robust classes, such as labeling vowels as front/back, can reduce the number of word candidates. In addition, broad classes for glottalization and for aspiration, which frequently occurs at the end of a sentence, especially after a sentence-final /r/, would be helpful since these sounds were not strongly associated with any of the six broad classes.

The boundaries assigned to segments were sometimes offset in time from the hand transcription, resulting in "extra" substitutions. Errors attributable to this offset are more evident in short broad phonetic segments, such as a short voiced obstruent or a word initial nasal. The offset is in part due to arbitrarily dividing the transition regions between the adjacent segments. By reducing the number of offset errors by the broad phonetic classifier, the correct word scores would be more closely concentrated near a probability of 1, and a larger percentage of the word hypotheses could be pruned. Two options for "reducing" the offset error statistics and therefore improving the performance are to: 1) try to find more accurate boundaries in the broad phonetic classifier or 2) compute performance statistics by using a *string matching* algorithm (e.g., Levenshtein distance in Sankoff and Kruskal, 1983) to match the sequence of broad phonetic labels produced by the system witb the hand labeled phonetic transcription. In the second option, an assumption is made that the segment boundaries will be adjusted in verification. This option is preferred because an offset in the location of segment boundaries would not be counted as an insertion or deletion. In addition, the location of more accurate segment boundaries is postponed until the identity of the hypothesized phone segments is known.

5.9.2 Lexical Component

In the lexical component, the use of durational, path, and allophonic constraints needs to be further investigated. The number of word candidates remaining after each constraint is applied should be evaluated over a variety of sequential constraint thresholds. Durational constraints were used to specify when a broad phonetic segment could represent only one or only two phones. These constraints were not

used for individual phones because the duration of a phone can vary greatly from utterance to utterance. A reference duration, which could be derived, for example, from knowledge of the number of digits in a digit string, is needed to effectively apply durational constraints at the phone level.

Methods for handling noise in tbe speech signal, such as lip smacks, need to be incorporated into the application of path constraints. One possibility is to allow possible noise segments to be skipped in a path at a specified cost.

The rules describing allophonic constraints were determined empirically from examination of spectrograms. A more formalized method should be used for defining the rules. For example, the broad phonetic representation (produced by the system) for each pronunciation and context of a word could be statistically tabulated, and these statistics could be used to weight the score assigned to a word. The word scores could then be characterized, as was done in the application of sequential constraints, and a threshold set sucb that poorly scoring words are removed from further consideration.

5.9.3 Verifier

With a knowledge-based system, more knowledge can always be added. In addition to development of additional detailed acoustic features for verification of phone hypotheses, a method for optimizing the set of features should prove to be valuable in removing features which do not contribute much information.

Allophones of a phoneme were sometimes grouped together in the training statistics. For example, the acoustic realizations of /r/ are different when in prevocalic, postvocalic, and intervocalic position. The current implementation of the verifier primarily used the feature of /r/-Possibility, which was computed over the middle 50% of a phone, to estimate how well a phone is realized as an /r/. This feature favors intervocalic /r/'s, which are most "/r/-like" in the center of a phone, over prevocalic and postvocalic /r/'s. By treating each allophone of /r/ as a different phone, better results may be obtained.

A weakness in the acoustic-phonetic approach as we implemented it is that the regions of speech used for training and testing were different. Thus, the values obtained in testing may be different. Hence, a method such that training and testing are performed on the same regions of speech should be developed.

A measure of spectral concentration was used to provide rough information about formants, rather than using a formant tracker. When current formant trackers fail, they may make gross errors. The development of a reliable formant tracker or one which indicated the reliability of the computed formant values would be useful in verification. The spectral concentration measure was found to be satisfactory for the limited digit vocabulary; however, a more accurate measure of formant location is needed for other vocabularies.

This research used the digit vocabulary as the basis for exploring how speech constraints can be applied to speech recognition. Pursuing this study for other vocabularies will require development of a more sophisticated control strategy in the verifier, perhaps involving the combination of expert system techniques with multivariate statistics. Coarticulation was largely ignored by defining features over the regions of a phone least affected by coarticulation. With a larger vocabulary, coarticulation will be more important in the identification of phones. One way in which coarticulation effects could be included is by developing a more sophisticated control strategy which can reason about conflicting cues. A control strategy could also be used to handle the varying number of phones in a segment so that normalization by duration or number of segments, both of which have faults, is not necessary.

The verifier was evaluated using incremental simulation. This technique isolated errors due to earlier components from errors due to the verifier, thus allowing the use of acoustic-phonetic features for verification of phones and words to be explored. Before the verifier can effectively use the word lattice produced by the

lexical component, methods need to be developed so that the verifier can handle initial broad class segmentation errors. These methods could include refinement of the broad class boundaries using information about the phonetic transcription of each hypothesized word before verification is performed.

5.9.4 Extensions to Other Tasks

The digit vocabulary formed a well-constrained task in which the utility of speech constraints could be studied in each component of the recognition model. One way to extend this thesis is to modify the knowledge used by the system to include other vocabularies. To extend the system to other vocabularies, the lexicon must be modified to include words from the new vocabulary. In addition, the broad phonetic classifier statistics need to be updated to handle any new phones and phone pair sequences.

A vocabulary could be chosen to illustrate a particular component of the recognition model. For example, a vocabulary composed primarily of dissimilar polysyllabic words could be used to study the performance of lexical access. Ideally, most of the word hypotheses produced by the lexical access component for this vocabulary should be uniquely specified. In contrast, a vocabulary composed of similar words could be chosen to study the use of acoustic features in the verification component.

In summary, this thesis has explored the viability of an acoustic-phonetic recognition model for continuous speech. Using the digits as a case study, the model was shown to be a viable approach to speech recognition which is potentially speaker-independent. We believe that further research should be pursued in order to fully develop this approach.

Appendix A

The Digit Corpus

The digit corpus was composed of 22 seven-digit strings (Corpus A) and 100 random-length digit strings. The random-length digit strings were divided into two subsets, Corpus B and Corpus C. Corpus B was included in the training set so that the system could be trained on random-length digit strings as well as on 7-digit strings. Corpus C was used for evaluating the system components.

The 7-digit strings were defined such that each sequence pair of digits, not including the pair formed by the third and fourth digits, was uniformly represented. The pair formed by the third and fourth digits was not considered in anticipation of people pausing between the third and fourth digit, as when saying a telephone number. However, the speakers were not told to pause, but instead were instructed to say the digit strings as naturally as possible. In addition, the representation of each digit in the string and representation at sentence initial and sentence final position was balanced over the corpus.

The length of the random-length strings was uniformly distributed from one to seven digits. The strings were generated using a random number generator to select the length of the string and the numbers composing the string. The numbers in the string were evaluated for uniform representation of pair sequences. A few strings were edited to insure that each sequence was represented at least once and that

each digit was represented at least once in isolation.

All utterances were orthographically transcribed. In addition, a subset of these utterances was also phonetically transcribed. These transcriptions were manually time-aligned to the speech waveform and with each other using the Spire facility available on the MIT Lisp Machine Workstations (Shipman, 1982; Cyphers, 1985). The position of a segment boundary was determined from observation of the speech spectrogram, the expanded speech waveform, the short-time spectra, and if necessary, by listening to a region of speech.

A few rules were used for transcribing ambiguous cases:

1. Because there usually are no clear boundaries between a vowel and liquid/glide, these boundaries were marked by assigning a fixed proportion of the vocalic region to each label unless the proportion definitely looked incorrect. Non-intervocalic liquid/glides are assigned $\frac{1}{3}$ of the vowel-liquid/glide region. Intervocalic liquid/glides were assigned $\frac{1}{3}$ of the region from the beginning of the preceding vowel to the midpoint of the liquid/glide, plus $\frac{1}{3}$ of the region from the midpoint of the liquid/glide to end of the following vowel.

2. Vowel to vowel boundaries were sometimes difficult to establish when the formants moved smoothly and no glottalization occurred. These boundaries were marked at the middle of the transition between two vowels.

3. When two phones geminate across a word boundary, as in "six seven" or "one nine," the geminate segment was split evenly between the two words unless clear cues to a boundary were evident.

4. When only one release was present in the sequence "eight two," then the closure is assigned to the "eight" and the release is assigned to the "two".

5. Glottalization occurring between two vowels at a word boundary was assigned to the second word.

The utterances were recorded using a Sennheiser noise canceling microphone in a "quiet" room. Corpus A is recordings of four male and three female speakers. Corpus B and Corpus C are recordings of five male and five female speakers. Two of the males and three of the females were the same for the two corpora.

Corpus A			
0315796	1807227	2964898	3674219
4583510	6093882	8240103	9253394
7620085	5471181	6327812	5043023
6861994	7352395	4159706	2869497
8436401	7698316	6077153	5532742
4615931	3214200	2468135	

Corpus B			
43039	90678	88361	56
981357	066	574	69
54	6394829	8846	15
65	019	27581	24
8517	516	733658	7
06	85031	5	38872
307	3		

Corpus C			
9350311	6521325	54434	9
595	89	37	851
1496070	65298	35569	816
5903	432678	91	374
3	547750	0407917	6844
2	6	9005010	61528
57345	15	237	6
2	0	2645	1
4	94	006097	081
4000	52	958399	19120
903	7	50717	1
32	542621	020157	559069
598	260539	686766	40853
305	98959	8762	5
305	0522945	49162	8158054
722	8	5659020	138940
7792	18	7326	541135
780933	13003	66	785753
80	69900		

The utterances in the corpus were divided into four categories:

1. training utterances by training speakers

2. new utterances by training speakers

3. training utterances by new speakers

4. new utterances by new speakers

The numbers followed by a "u" are orthographically but not phonetically transcribed. The utterances in category one form the training set for each of the system components. Thus, the components were trained on random-length digit strings and 7-digit strings spoken by three male and three female speakers.

Subsets of Evaluation Corpus

Speaker	male/female	Corpus Subset		
		A	B	C
jrg	m	1	1	2
mar	m	1	1	2
sch	m	1		
ama	f	1	1	2
cab	f	1	1	2u
chs	f	1	1	2
rhk	m		3	4
wpd	m		3	4
lsp	m		3u	4u
lfl	f		3	4
lsl	f	3	3u	4u

Appendix B

Sample Production Rules

This is an example of a production rule for hypothesizing the phone-like class of strong-fricative-like from acoustic features:

```
(defrule (strong-fric-like1 *initial-rules*)
  (if (and hi-zc
           hi-hfe
           (at-most-one-of vocalic-1 hi-lfe))
      (assert strong-fric-like)))
```

This rule states that if (1) the zero-crossing-rate is high, and (2) the high-frequency-energy is high, and (3) at most one of the indicators of high low-frequency-energy is on, then assert that a strong fricative may be present in the region.

This is an example of a production rule for hypothesizing the phone class of strong-fricative from the phone-like classes:

```
(defrule (strong-fric1 *class-rules*)
  (context-of (((anything)
                ((with-duration > 10 strong-fric-like))
                (anything))
               (if (and (max-greater phfe -65)
                        (max-greater pte -55))
                   (assert strong-fricative)
                   (and (assert strong-fricative)
                        (assert weak-fricative))))))
```

This rule states that if (1) a strong-fric-like segment was hypothesized which has a duration cf at least 10 msec, (2) the segment is preceded hy anything, and (3) the segment is followed hy anything, then a strong-fricative and possihly a weak-fricative is asserted. If the maximum value of high-frequency-energy in the region is greater than -65 dB, and the maximum value of total-energy in the region is greater than -55 dB (the threshold values were determined from a statistical characterisation of the phone classes), then only a strong-fricative is asserted; otherwise, both a strong-fricative and a weak-fricative are asserted. Note that dnrational and contextual constraints can he specified, as well as additional aconstic features.

Appendix C

Insertions and Deletions

Insertions

The following tahle lists the insertion errors in the ontput of the hroad phonetic classifier which contribute at least 1% of the total nnmber of insertion errors. Note that the errors are reasonable. For example, the two predominant errors are insertion of silence in the laheling of tho weak fricatives /f/ and /θ/. As another example, the vowels with offglides form a group in which the offglide portion of the vowel is labeled as a sonorant.

count	broad labels	phone label
25	(SILENCE WEAK-FRIC)	f
25	(SILENCE WEAK-FRIC)	θ
22	(VOWEL SONORANT)	u
19	(SONORANT VOWEL)	ɔ
18	(VOWEL SONORANT)	a^y
17	(VOWEL SONORANT)	o^w
17	(VOWEL SONORANT)	ɪ^y
15	(VOWEL SONORANT)	e^y
14	(STRONG-FRIC WEAK-FRIC)	s
13	(SONORANT SILENCE)	n
12	(WEAK-FRIC SILENCE)	f
11	(VOWEL SONORANT)	ü
10	(WEAK-FRIC STRONG-FRIC)	s
8	(WEAK-FRIC SILENCE)	t
8	(VOWEL SONORANT)	n
7	(VOWEL SONORANT)	ɔ
7	(SILENCE STRONG-FRIC)	s
6	(VOWEL SILENCE)	n
6	(VOWEL SONORANT)	r
6	(SILENCE WEAK-FRIC)	v
6	(VOWEL SILENCE)	v
6	(STRONG-FRIC WEAK-FRIC)	t
5	(WEAK-FRIC SONORANT)	θ
5	(SONORANT VOWEL)	r
5	(VOWEL V)	v
5	(SONORANT VOWEL)	ɨ

Deletions

The following table lists the deletion errors in the output of the broad phonetic classifier which contribute at least 1% of the total number of deletiou errors. Note that the errors are reasonable. For example, the system was not designed to identify /r/; /r/ was considered to be a vowel. We see that the first four predominant errors are deletion of /r/ when adjacent to a vowel. In addition, the system did not try to locate [k] separately from its closure or the following fricative. Thus the errors of calling [ks] a strong fricative and [k^{c}k] silence are also reasonable errors.

count	broad label	phone labels
129	VOWEL	r ɾ
122	VOWEL	ɪ r
117	VOWEL	r o^w
90	VOWEL	ɔ r
60	STRONG-FRIC	k s
48	SILENCE	k^c k
32	VOWEL	ɛ v
26	VOWEL	w ʌ
25	VOWEL	v ə
24	VOWEL	ə n
21	SONORANT	ə n
20	VOWEL	ʌ n
18	VOWEL	a^y v
18	VOWEL	ʔ e^y
15	VOWEL	a^y n
14	SONORANT	v ə
14	VOWEL	r ə
13	WEAK-FRIC	o θ
13	SILENCE	t^c t
12	VOWEL	ə r
12	STRONG-FRIC	s z
12	WEAK-FRIC	t f
12	SONORANT	n w

Glossary

Some of the terms in this document have different meanings as used by different people. This Glossary is an attempt to clarify the intended meaning of some of these words as used in this thesis.

broad phonetic class: A set of phones which have common acoustic characteristics that can be robustly identified.

broad phonetic level: A description of the speech signal in which the signal is represented as a sequence of segments and each segment is labeled as a broad phonetic class.

constraint: A restriction on the search space. In this thesis, speech knowledge is formulated into constraints which are used to identify poor word hypothesis and rule them out from further consideration

front end: In reference to the acoustic-phonetic recognition model, this is the broad phonetic classifier.

feature/cue: A representation of a region of a parameter which attempts to capture a salient characteristic of the parameter and which may be related to one or more speech sounds.

low level speech knowledge: Characteristics about speech derived from the acoustic signal.

parameter: A set of values directly derived from the speech signal. These values can be used to characterize the speech signal on a sample by sample basis (forming a feature vector) or can be characterized into features. The difference between a parameter and a cue may sometimes be ambiguous.

segment: A region of speech. This region may be associated with a variety of speech units, such as a phone or broad phonetic class.

segmentation string: Representation of the speech signal by a sequence of discrete units, each of which is associated with a label. In the output produced by the broad phonetic classifier, the labels are one of six broad phonetic classes.

sonorant: This term refers to the class of sonorant consonants, in contrast to the distinctive feature used by linguists.

References

Bahl, L.R., A.G. Cole, F. Jelinek., R.L. Mercer, A. Nadas, D. Nahamoo, and M.A. Picheny. "Recognition of Isolated-Word Sentences From a 5000-Word Vocabulary Office Correspondence Task," *Proceedings of the IEEE Internat. Conf. on Acoustics, Speech, and Signal Process.*, pp. 1065-1067, 1933.

Bush, M.A. and G.E. Kopec. "Network-Based Connected Digit Recognition Using Vector Quantization," *Proceedings of the IEEE Internat. Conf. on Acoustics, Speech, and Signal Process.*, pp. 1197-1200, 1985.

Buzo, A., A. Gray, R. Gray, and J. Markel. "Speech Coding Based Upon Vector Quantization," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. ASSP-28, no. 5, pp. 562-574, 1980.

Chomsky, N., and M. Halle. *The Sound Pattern of English*, Harper and Row: New York, 1968.

Cole, R.A., A.I. Rudnicky, V.W. Zue, and D.R. Reddy. "Speech as Patterns on Paper," in *Perception and Production of Fluent Speech*, Edited by R.A. Cole, Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.

Cole, R.A., M.S. Phillips, S.M. Brill, P. Specker, and A.P. Pilant. "Speaker-independent Recognition of English letters," Paper presented at the 104th meeting of the Acoustical Society of America, Orlando, FL, 1982.

Cole, R.A., R.M. Stern, and M.J. Lasry. "Performing Fine Phonetic Distinctions: Templates vs. Features," in *Invariance and Variability of Speech Processes*, Edited by J. Perkell and D. Klatt, Hillsdale, NJ: Lawrence Erlbaum Associates, forthcoming.

Cook, C.C., and R.M. Schwartz. "Advanced Acoustic Techniques in Automatic Speech Understanding," *Proceedings of the IEEE Internat. Conf. on Acoustics, Speech, and Signal Process.*, pp. 663-666, 1977.

Cyphers, D.S. "Spire: A Speech Research Tool," S.M. thesis, Massachusetts Institute of Technology, 1985.

Doddington, G.R. and T.B. Scholk. "Speech Recognition: Turning Theory to Practice," *IEEE Spectrum*, pp. 26-32, Sept. 1981.

Duda, R.O. and P.E. Hart. *Pattern Classification and Scene Analysis*, New York: John Wiley and Sons, 1973.

Erman, L.D. and V.R. Lesser. "The Hearsay II Speech Understanding Syrtem: A Tutorial," in *Trends in Speech Recognition*, Edited by W.A. Lea, Englewood Cliffs: Prentice-Hall, Inc., 1980.

Fant, G. "Distinctive Features and Phonetic Dimensions," *Speech Transmission Laboratory Quarterly Progress and Status Report*, STL-QPSR 2-3, pp. 1-18, 1969.

Fant, G. *Acoustic Theory of Speech Production*, The Hague: Mouton, 1970.

Fujimura, O. "Analysis of Nasal Consonants," *Journal of the Acoustical Society of America*, Vol. 34, No. 12, pp. 1865-1875, 1962.

Fujimura, O. "Syllable as a Unit of Speech Recognition," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. ASSP-23, no. 1, pp. 82-87, 1975.

Heffner, R-M.S. *General Phonetics*, Madison: The University of Wisconsin Press, 1950.

House, A.S., and E.P. Neuburg, " Toward Automatic Identification of the Language of an Utterance. I. Preliminary Methodological Considerations," *Journal of the Acoustical Society of America*, Vol. 62, No. 3, pp. 708-713, 1977.

Hyde, S.R. "Automatic Speech Recognition: A Critical Survey and Discussion of the Literature," in *Human communication: A Unified View*, Edited by E.E. David, Jr. and P.B. Denes, New York: McGraw Hill, pp. 399-438, 1972.

Itakura, F. "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. ASSP-23, pp. 67-72, 1975.

Jakobson, R., C.G.M. Fant, and M. Halle. *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates*, Cambridge, Mass.: The MIT Press, 1952.

Jelinek, F., L.R. Bahl, and R.L. Mercer. "Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech" *IEEE Trans. Infor. Theory*, vol. IT-21, pp. 250-256, May 1975.

Jelinek, F. "Continuous Speech Recognition by Statistical Methods," *Proceedings of the IEEE*, vol. 64, No. 4, pp. 532-556, 1976.

Jelinek, F., R.L. Mercer, and L.R. Bahl. "Continuous Speech Recognition: Statistical Methods," unpublished paper, 1980.

Jelinek, F. "Self-Organized Continuous Speech Recognition," *Proceedings of the NATO Advanced Summer Inst. Auto. Speech Analysis and Recognition*, Bonas, France, 1981.

Kaplan, G. "Words into action:I," *IEEE Spectrum*, pp. 22-26, June 1980.

Kato, Y. "Words into action III: a commercial system," *IEEE Spectrum*, p. 29, June 1980.

Klatt, D.H. "Review of the ARPA Speech Understanding Project, *Journal of the Acoustical Society of America*, Vol. 62, No. 6, pp. 1345-1366, 1977.

Lea, W.A., ed. *Trends in Speech Recognition*, Prentice Hall Inc., Englewood Cliffs, New Jersey, 1980.

Lesser, V.R., R.D. Fennell, L.D. Erman, and D.R. Reddy. "Organisation of the Hearsay II Speech Understanding System," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. ASSP-23, pp.11-24, 1975.

Lowerre, B. "Dynamic Speaker Adaptation in the Harpy Speech Recognition System," *Proceedings of the IEEE Internat. Conf. on Acoustics, Speech, and Signal Process.*, pp. 788-790, 1977.

Lowerre, B., and D.R. Reddy. "The Harpy Speech Understanding System," in *Trends in Speech Recognition*, Edited by W.A. Lea, Englewood Cliffs: Prentice-Hall, Inc., 1980.

Mermelstein, P. "A Phonetic-Context Controlled Strategy for Segmentation and Phonetic Labeling of Speech," it IEEE Trans. Acoustics, Speech, and Signal Process., vol. ASSP-23, pp. 79-82, 1975.

Myers, C.S., and L.R. Rabiner. "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. ASSP-29, pp. 284-297, 1981a.

Myers, C.S., and L R. Rabiner. "Connected Digit Recognition Using a Level-Building DTW Algorithm," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. ASSP-29, pp. 351-363, 1981b.

Oppenheim, A.V., and R.W. Schafer, "Homomorphic Analysis of Speech," *IEEE Trans. on Audio and Electroacoustics*, Vol. AU-16, No. 2, pp. 221-226, 1968.

Otten, K.W. "Approaches to the Machine Recognition of Conversational Speech," in *Advances in Computers*, Edited by F.L. Alt, M. Ribinoff, and M.C. Yovits, New York: Academic Press, Vol. 11, pp. 127-163, 1971.

Pan, K.C., F.K. Soong, L.R. Rabiner, and A.F. Bergh. "An Efficient Vector-Quantization Preprocessor for Speaker Independent Isolated Word Recognition," *Proceedings of the IEEE Internat. Conf. on Acoustics, Speech, and Signal Process.*, pp. 874-877, 1985.

Pavlidis, T. and S.L. Horowitz. "Segmentation of Plane Curves," *IEEE Trans. on Computers*, vol. C-23, no. 8, pp. 860-870, 1974.

Peterson, G.E. and H.L. Barney. "Control Methods Used in a Study of the Vowels," *Journal of the Acoustical Society of America*, 24(2) pp. 175-184, 1952.

Rabiner, L. "On Creating Reference Templates for Speaker-Independent Recognition of Isolated Words," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. ASSP-26, no. 1, pp. 34-42, 1978.

Rosenburg, A.E., L.R. Rabiner, J.G. Wilpon, and D. Kahn. "Demisyllable-Based Isolated Word Recognition System" *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. ASSP-31, no. 3, pp. 713-726, 1983.

Sakoe, H. "Two-Level DP-Matching–A Dynamic Programming-Based Pattern Matching Algorithm for Connected Word Recognition," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. ASSP-27, pp. 588-595, 1979.

Sankoff, D. and J. Kruskal (eds.), *Time Warps, Strong Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Reading, Mass.: Addison-Wesley, 1983.

Scagliola, C. and L. Marmi. "Continuous Speech Recognition via Diphone Spotting. A Preliminary Implementation," *Proceedings of IEEE Internat. Conf. on Acoustics, Speech, and Signal Process.*, pp. 2008-2011, 1982.

Schwartz, R., Y. Chow, O. Kimball, S. Roucos, M. Krasner, and J. Makhoul. "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech," *Proceedings of IEEE Internat. Conf. on Acoustics, Speech, and Signal Process.*, pp. 1205-1208, 1985.

Schwartz, R.M. and V.W. Zue. "Acoustic-Phonetic Recognition in BBN SPEECHLIS," *Proceedings of IEEE Internat. Conf. on Acoustics, Speech, and Signal Process.*, pp. 21-24, 1976.

Shipman, D.S. and V.W. Zue. "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems," *Proceedings of IEEE Internat. Conf. on Acoustics, Speech, and Signal Process.*, pp. 546-549, 1982.

Shipman, D. "Development of Speech Research Software on the MIT Lisp machine," Paper presented at the 103rd meeting of the Acoustical Society of America, Chicago, Il., April 1982.

Sigurd, B. "Phonotactic Aspects of the Linguistic Expression," in *Manual of Phonetics*, 2nd ed., Edited by B. Malmber, Amsterdam: North Holland Publishing Co., 1970.

Winston, P.H. *Artificial Intelligence*, 2nd ed., Reading, MA: Addison-Wesley, 1984.

Wolf, J.J. and W.A. Woods. "The HWIM Speech Understanding System," in *Trends in Speech Recognition*, Edited by W.A. Lea, Englewood Cliffs: Prentice-Hall, Inc., 1980.

Woods, W., M. Bates, G. Brown, B. Bruce, C. Cook, J. Klovstad, J. Makhoul, B. Nash-Webber, R. Schwartz, J. Wolf, and V. Zue. "Speech Understanding Systems: Final Technical Progress Report," Bolt Beranek and Newman, Inc, Report No. 3438, Vol. II, Cambridge, Mass., 1976.

Woods, W.A. "Optimal Search Strategies for Speech Understanding Control," in *Readings in Artificial Intelligence*, Edited by B.L. Webber and N.J. Nilsson, Palo Alto, Calif.: Tioga Publishing Company, 1981.

Nasal Consonants and Nasalized Vowels: An Acoustic Study and Recognition Experiment

by

James Robert Glass
B.Eng., Carleton University
(1982)

Submitted in Partial Fulfillment
of the Requirements for the
Degrees of

Master of Science

and

Electrical Engineer

at the

Massachusetts Institute of Technology

December 1984

Signature of Author ....James R. Glass.......
Department of Electrical Engineering and Computer Science
December 21, 1984

Certified by ....V. W. Zue.......
Victor W. Zue
Thesis Supervisor

Accepted by ....
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

Nasal Consonants and Nasalized Vowels: An Acoustic Study and Recognition Experiment

by

James Robert Glass

Submitted to the Department of Electrical Engineering and Computer Science on December 21, 1984 in partial fulfillment of the requirements for the degrees of Master of Science and Electrical Engineer.

Abstract

This thesis is concerned with the acoustic analysis of nasal consonants and nasalized vowels, and the design, implementation, and evaluation of a set of algorithms to detect nasal consonants and nasalized vowels from the speech waveform. The acoustic study uses a database consisting of over 1200 words, excised from continuous speech, and recorded from six speakers, three male and three female. All of the recorded words were digitized, and their pbonetic transcriptions were aligned with the speech waveform. Using the Spire and SpireX speech analysis tools, acoustic features common to all nasal consonants and nasalized vowels were determined. For nasal consonants these included the presence of a low frequency resonance in the short-time spectra, centered between 200 and 350 Hz, the global, and local strength of this peak, and a measure of spectral stability in the low frequency regions. For nasalized vowels these included the presence of an extra resonance in the short-time spectra, the relative amplitude of this peak to that of the first formant, and a measure of the hroadness of the spectral peak in the first formant region.

The nasal consonant detection algorithms were designed to discriminate hetween nasal consonants and impostor sounds such as liquids, glides, or voice bars. The nasalized vowel algorithms were designed to discriminate vowels adjacent to a nasal consonant from vowels in other contexts. In each case, a log likelihood decision strategy, using robust measures established in the acoustic analysis, was employed. The detection systems were evaluated on the datahase by training on the speech of five speakers, and then testing on the tokens of the final speaker. The results indicate that a nasal consonant can he detected 88% of the time, while a vowel adjacent to a nasal consonant can he identified 74% of the time.

Thesis Advisor: Victor W. Zue
Title: Assistant Professor of Electrical Engineering and Computer Science

Acknowledgments

I am extemely grateful to my advisor, Victor Zue, for giving me the opp unity to learn from his expertise in speech analysis and speech recognition. I thank him for his guidance, especially dnring the difficnlt times, and for his advice dnring the preparation of tbis tbesis.

I also thank the members of Znes Crew, and the Speech Communication Group, for many belpful snggestions and discussions, and for creating a stimnlating and enjoyahle environment in which to conduct research. I thank Mark Randolph in particnlar, for assisting me with the use of his hinary tree classifier, and Scott Cyphers, who spent countless hours maintaining the laboratory Lisp Machines, and improving the Spire facility.

I thank Lori Sorayama, for carefully reviewing drafts of this thesis, and for ber support and encouragement for as long as I have known her at MIT.

Finally, I thank my parents, and my brother and sisters, for their continuing love, and encouragement, and for showing me the value of an edncation.

Tbis researcb was supported hy the Natural Sciences and Engineering Researcb Council of Canada, and by tbe Office of Naval Research.

Contents

Contents

Chapter 1

Introduction

1.1 Machine Recognition of Speech

The topic of automatic speech recognition has intrigued scientists and engineers for many years. Apart from the brief period of large scale effort in continuous speech recognition witnessed during the ARPA project [28], the majority of research in this field has been directed towards isolated word recognition. This is particularly true of the recent past, where the primary focus of attention has been on the development of small-vocabulary, speaker-dependent, isolated-word recognition systems. These systems tend to be based on general pattern matching techniques and incorporate little speech specific knowledge [25], [34].

Although general pattern matching algorithms excel within their limited problem space, the extension of these techniques to more difficult tasks involving multiple speakers, large vocabularies, or continuous speech have largely been met with limited success. These results have caused many researchers to believe that large recognition systems would be more successful, if they incorporated a better understanding of speech sounds. This belief is reinforced, at least in part, by a series of spectrogram reading experiments by Cole et al, which indicated that the acoustic signal is rich in phonetic information [7]. These experiments revealed that a trained subject, using explicit acoustic-phonetic rules, could phonetically

transcribe unknown sentences from speech spectrograms with an accuracy of 85%. This result suggests that automatic phonetic recognition performances have the potential to be substantially better than are presently reported [28].

One of the most important factors leading to this benchmark performance in spectrogram reading was an improved understanding of the acoustic characteristics of fluent speech. Although there has been a significant amount of research over the last forty years on the acoustic properties of speech sounds, little attention has been given to the acoustic characteristics of speech sounds in continuous speech. Over the last decade this has slowly been changing. As Zue has illustrated, we now have a much better understanding of the properties of speech sounds in different phonetic environments [78]. However, there is still a need for basic research directed towards the *quantification* of the acoustic characteristics of speech sounds.

The research in this thesis is motivated with this requirement in mind. The primary objective of this work is to characterize, and quantify, the acoustic properties of nasal consonants and nasalized vowels in American English. Nasal consonants were chosen because they appear to cause difficulty for some speech recognition systems, yet have not been studied as extensively as many other speech sounds. Nasalized vowels were included because of clear indications that they provide important acoustic information about the presence of a nasal consonant.

Once the characteristics of nasal consonants and nasalized vowels are quantified, automatic detection systems, which incorporate robust acoustic measures of nasality, are designed for use in a speaker-independent, continuous-speech environment. Evaluation of these systems provides an indication of their potential for use in speech recognition.

1.2 Acoustic Studies of Speech

1.2.1 The Nature of Speech Sounds

All languages appear to consist of a finite number of distinguishable, mutually exclusive sounds which are concatenated together in time to produce speech. These basic linguistic units are called *phonemes*, and possess unique articulatory and acoustic characteristics [14]. In American English, there are approximately 42 phonemes, which include vowels, semivowels, and consonants [13].

It has long been proposed that there are underlying invariant acoustic properties for all phonemes, which allow an utterance to be decoded from the acoustic signal [26]. However, there are many factors which can influence the observed acoustic pattern of phonemes, and therefore complicate a study of their properties. These factors include:

- *Contextual differences.* When phonemes are connected together to form larger linguistic units, the acoustic characteristics of a given phoneme are modified by the immediate phonetic environment. Occasionally, a speaker can distort the acoustic properties so severely that the phoneme may not be identified, despite a knowledge of the phonetic environment [76]. These distortions are possible because, in addition to acoustic-phonetic knowledge, listeners are able to apply syntactic, semantic, phonetactic, and phonological constraints to help recognize an utterance.

- *Inter-speaker differences.* The acoustic characteristics of speech sounds depend upon the physiological structure of the vocal apparatus which varies from speaker to speaker. In particular, there can be large acoustical differences in the speech of men, women, and children.

- *Intra-speaker differences.* The same speaker can pronounce an utterance differently on separate occasions for many reasons including sickness, mood,

audience (e.g. child versus adult), stress patterns on the word or phrase, and transmission environment.

In order to compensate for these factors, many studies, including this one, base their analysis on a carefully designed database. A discussion of the motivation for utilizing databases, and a description of the particular database used in this analysis, are presented later in more detail.

1.2.2 Production of Nasal Sounds

The basic production mechanisms of nasal consonants and nasalized vowels have been studied extensively and are well understood [13], [14]. Nasal consonants are considered to be voiced, since during their production the vocal tract is excited by vocal fold vibration. Nasal consonants are produced by lowering the velum so that air flows through the nasal tract and is radiated at the nostrils (figure 1.1 shows a cross-section of the human vocal apparatus). The closed oral cavity and the sinuses of the nose form shunting cavities to the main path (pharynx and nasal tract) which substantially influences the resulting radiated sound. Figure 1.2 illustrates typical vocal tract configurations for /m/, /n/, and /ŋ/, the three nasal consonants produced in American English. Note that the main difference between the three consonants is the location of the constriction formed with the tongue. Figure 1.3 contains spectrograms of the words *simmer*, *sinner*, and *singer*.

Nasalized vowels are produced in a similar manner to nasal consonants, with the exception being that the oral cavity is not blocked, thereby allowing air to flow through *both* the nasal and oral cavities.

In many languages, including American English, nasal consonants can have a profound effect on neighboring vowels. Following the release of a nasal consonant, the initial portion of a following vowel will be nasalized during the time interval that the velum is closing. The same holds true for the final portion of a vowel

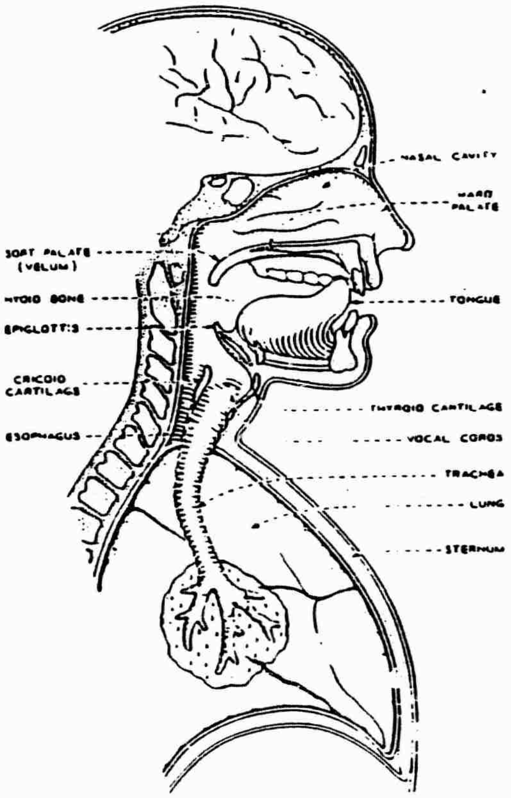


Figure 1.1: The Human Vocal Apparatus (from Flanagan)

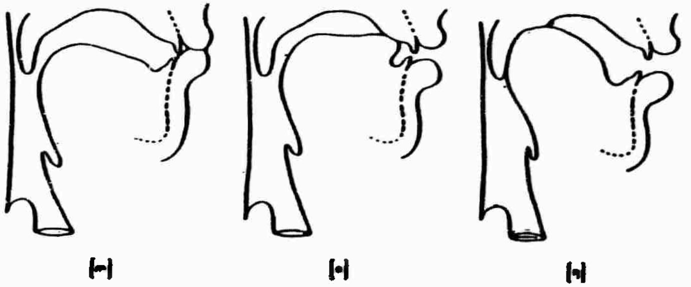


Figure 1.2: The Vocal Tract Configurations of the Nasal Consonants

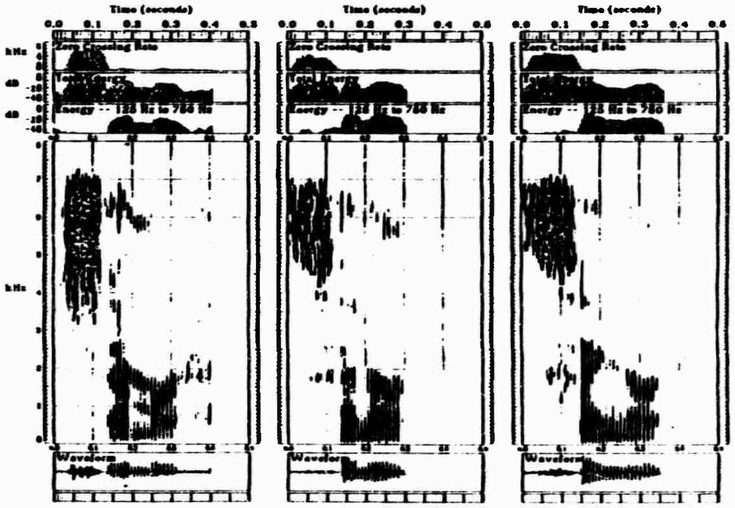


Figure 1.3: Spectrograms of the words *simmer*, *sinner*, and *singer*

preceding a nasal consonant [21]. The amount of coarticulated nasalization depends upon the particular language and dialect. Since anticipatory nasalization is common in American English [27], a sequence of a vowel plus a nasal consonant (VN) may, in many situations, be pronounced as a simple nasalized vowel, or a nasalized vowel plus a short, residual nasal murmur. This is especially true of vowel-nasal-consonant (VNC) sequences where the consonant is a voiceless stop, as in the words *cump*, *bent*, or *bunk* [14]. In these cases, nasalization of the preceding vowel may provide the major acoustic difference between these words, and the corresponding pair words *cup*, *bet*, and *buck*.

In American English, nasalized vowels are not distinguished phonemically from non-nasalized vowels. Thus, speakers have the freedom to nasalize vowels at will, *independent of the presence or absence of a nasal consonant*. For this reason, it is important not to assume that the presence of a nasalized vowel will always indicate the presence of a nasal consonant as well. This research determines if the *relative* degree of vowel nasalization is a more robust indication of the presence, or absence, of a nasal consonant.

In addition to the potential benefit to nasal consonant detection, an understanding of the acoustic characteristics of nasalized vowels would be very useful for many speech analysis tools such as formant trackers, which have traditionally had difficulty with nasalized vowels. Knowledge of the nasalized portions of an utterance would allow formant trackers to employ different, and more successful, strategies in these regions.

1.2.3 Previous Studies of Nasal Sounds

There is a vast amount of literature spanning over twenty-five years which involves the analysis, synthesis, perception, and recognition of nasal consonants and nasalized vowels. The following sections attempt to provide a brief summary of

some of this work in order to put the acoustic study of this research into better perspective.

Analysis and Synthesis Studies

There has been a large amount of work which has studied the acoustic characteristics of nasal consonants and nasalized vowels. Much of this research has involved the use of synthetic speech. The following paragraphs summarize some of past work on nasal consonants.

- Using an analog vocal tract model, House found that synthetic nasal consonants were characterized by a predominance of low frequency energy, low overall level compared to a vowel, and a spectral prominence near 1000 Hz [24].
- Fujimura reported several studies of the acoustic characteristics of nasal consonants [15], [16]. He found that the nasal murmur spectra is characterized by the existence of a very low first formant, located at about 300 Hz, which is well separated from the upper formant structure. He also noted that the formants were highly damped, and that there was a high density of formants compared to vowels. Further, he observed the presence of an antiformant, caused by the closure in the oral cavity, which varied in frequency with the place of articulation. In general, the antiformant could be found between 750 and 1250 Hz for /m/, between 1450 and 2200 Hz for /n/, and above 3000 Hz for /ŋ/. He noted however, that although the antiformant varies with the place of articulation, the overall spectral shape of nasal consonants are very similar in appearance.
- From sweep-tone measurements of the vocal tract, Fujimura and Lindqvist reported that the primary characteristics of nasal consonants were a marked, but not necessarily simple, low-frequency boost around 200 to 300 Hz [18].

They also noted a gross deviation from vowel spectral shapes, and a higher total energy compared to stops, both in the low-frequency boost, and in other frequency ranges. They also observed that the transfer function characteristics of the nasal consonants varied greatly from subject to subject.

- Based on evidence from sweep-tone data of the transfer function of the nasal tract, Lindqvist and Sundberg proposed that the complex pole zero patterns observed in nasals and nasalized vowels could be explained by the shunting effect of the sinus cavities [37].

The following paragraphs summarize past work with nasalized vowels.

- House and Stevens studied nasalized vowels with the use of an analog vocal tract synthesizer [23]. They found that the major characteristics of nasalization were a weakened, and broadened first formant, and an overall weaker vowel level than in non-nasalized vowels. They also observed additional weak spectral peaks which tended to fill in the valleys between formants.

- Hattori, Yamamoto, and Fujimura determined that the principal characteristics of nasalization were the presence of a dull resonance around 250 Hz, an antiresonance at about 500 Hz, and additional weak and diffuse components which filled in the valleys between formants [20].

- Fujimura and Lindqvist concluded that nasalization introduces nasal formants into the speech signal [18]. They found that each nasal formant was paired with an antiformant. Depending on the degree of coupling, the antiformant could be either close to the nasal formant, or a nearby oral formant. They also found that as nasalization increases, all formants shifted monotonically upwards.

- Maeda found that by including a model of the sinus cavities, he was able to synthesize a low resonance below the first formant [39]. The addition of this resonance was found to produce natural sounding nasalized vowels of all height.

All of these studies have contributed to the current understanding of the acoustic characteristics of nasal consonants and nasalized vowels. Despite these numerous acoustical studies however, the results are not always directly relevant to speech recognition. Reasons for this include the fact that the data has not been presented in sufficiently quantitative form, or has been presented in relative as opposed to absolute terms. More seriously for automatic speech recognition, some of the data has been obtained from displays where a human must make an interpretation to make a measurement. Finally, in many cases, the data has been obtained from restricted environments such as stressed, consonant-vowel (CV) syllables.

Further insights could be obtained by performing an analysis on a body of naturally spoken data. By providing a better understanding of the variability of the acoustic characteristics of nasal consonants and nasalized vowels in a natural speech environment, such a study would be valuable to scientists concerned with the automatic detection of these sounds.

Perceptual Studies

Much perceptual research has been devoted to studying the role of the nasal murmur and an adjacent vowel, in determining both the manner and place of articulation of the nasal consonant. This section summarizes the results of several of these studies.

- The work of Malécot, Nakata, Nord, Recasens, Kurowski and Blumstein, and Repp, has been concerned with the relative importance of the nasal murmur, and the formant transitions in adjacent vowels, to the identification of the place of the nasal consonant [43], [50], [53], [63], [31], [64]. The common conclusion was that formant transitions were the major cue to place for prevocalic nasals, while for post-vocalic nasals, the murmur was taken into account as well. The work of Kurowski and Blumstein, and Repp, found that the nasal murmur was more informative in utterance-initial position than did previous studies however.

- Malécot has also done perceptual work with homorganic nasal stop consonant clusters [44]. He found that the short nasal murmur played a very minor role in conveying the impression that a nasal was present. The nasalized vowel appeared to be the major cue to the presence of the nasal consonant.

- Mártony reported studies on synthetic nasal production which indicated that damping in the second formant region was very important for natural sounding NV tokens [46]. He also found that the bandwidth values for the nasal mormur of /m/ were much more vowel dependent than /n/ since in /m/ tends to be coarticulated with vowels more than /n/.[1]

- Ali et al reported an experiment indicating that subjects are able to predict the presence of a nasal consonant from the preceding vowel [1]. They hypothesized that listeners use the anticipatory nasalization feature, common for nasal production in English, to help lighten the phoneme processing load.

- Lintz and Sherman investigated the effect of different consonants on the perceived nasality of vowels in CVC tokens [38]. They found that low vowels were judged more nasal than high vowels, front vowels more nasal than back vowels. They also found that nasality is least severe for voiceless plosive environments, more severe for voiceless fricative and voiced plosive environments, and most severe for voiced fricative environments.

- Kawasaki found that vowels in NVN tokens were considered more nasalized when nasal murmor amplitudes were decreased relative to the vowel amplitude [27]. She also noted that playing the speech backwards made vowel nasalization much more apparent, and attributed this to the fact that listeners do not expect significant perservatory nasalization in English.

- Hawkins and Stevens have reported a perceptual study which indicates that the basic acoustic property of nasalization is a reduction in the degree of prominence of the first formant peak [21]. This reduction is realized by splitting or broadening the first formant spectral peak by creating an additional spectral peak nearby.

Perceptual studies have provided information about the role of different acoustic characteristics in establishing the property of nasality. From a speech recognition perspective, it would be useful to determine, based on acoustic information alone, the inherent recognizability of nasal consonants and nasalized vowels in American English. In other words, listeners would be allowed to use only their acoustic knowledge to decide on the feature nasal; syntactic, semantic and phonetactic information would, as much as possible, be eliminated. This test would provide two benefits. First, it provides an upper bound on automatic recognition performance in the same circumstances. Second, it provides a means of evaluating the perceptual relevance of a set of acoustic characteristics; a high correlation between acoustic measurements and perceptual score being the evaluation measure.

Recognition of Nasal Consonants

There have been several attempts at automatic recognition of nasal consonants:

- Gillmann reported his attempts at nasal identification for post-vocalic nasal consonants considering only the formants in the nasal murmur (by picking peaks of LPC spectra) [19]. He found that the formants did not change appreciably during the murmur, and that formant frequencies were fairly

[1] In American English the tongue has no distinctive function for /m/, unlike for /n/ or /ŋ/. Therefore, /m/ tends to be coarticulated with adjacent vowels much more than other nasal consonants [72]

stable for one speaker, although they varied from speaker to speaker. There were enough differeuces between nasal formant values for any one speaker that he was able to achieve 70% correct nasal identification using a simple least squares clustering procedure.

- Formant frequencies were also used to detect nasal consonants in the sonorant regions of the acoustic-phonetic analysis system developed by Weinstein et al. [73]. Nasal consonants were required to pass a duration constraint, as well as speaker dependent constraints on the formant values (such as low value of the first formant frequency, low ratio of second formant amplitude to first formant amplitude, and a higher ratio of third formant amplitude to first formant amplitude) to be accepted. They found that nasals were detected correctly about 80% of the time, with intervocalic nasals being detected much more reliably than non-intervocalic nasals. Prevocalic nasals were detected 80% of the time, while post-vocalic nasals were detected only 60% of the time due, in their opinion, to the reduction of the nasal murmur lengths in some environments. They noted that in these situations the adjacent vowel was quite often nasalized. About 15% of the detected pre- and post-vocalic nasals were the phonemes /l/, /w/, or /r/, and another 20% were false alarms (no segment present), caused by vowels with low first formant frequencies, such as /i/, /e/ and /u/.

- Using the hypothesis that nasal boundaries can be found at points of maximal spectral change, Mermelstein attempted a very ambitious project to detect nasal consonants in continuous speech [48]. He used four simple spectral measurements to classify the region adjacent to these transitions as either nasal or non-nasal. Using a multivariate statistical training procedure, he was able to obtain a 91% correct nasal/non-nasal decision rate on paragraphs spoken by two male speakers. Mermelstein also found that speaker dependent training was superior to speaker independent. He pointed out that the majority of errors confused nasals with weak fricatives and /l/ and /r/ before high vowels. He also pointed out that nasal segments were missed when they were shortened.

- Hess reported a 90% recognition rate for German nasals in continuous speech for a single speaker [22]. Dixson and Silverman reported a 94% recognition rate for nasals in continuous speech for one speaker [11].

- De Mori has reported work on discriminating intervocalic /n/ and /m/ in continuous speech [9]. Decision making was based on the value of the second formant at the beginning and end of the nasal consonant and the amplitude differences between the formants at the point during the nasal murmur where the second formant amplitude is minimal. Tested on four male speakers, the average error rate was 6% with the majority of error occurring in a front vowel environment.

There are two points which can be made about these studies. First, none of these efforts has reported testing their systems on a large number of speakers. The system was either designed to be speaker dependent, or was tested on very few speakers (all male). Clearly, the strong speaker dependent characteristics of nasal consonants present a challenge to any recognition system. In order to claim that a system is speaker independent, it is necessary to test it on a much larger number of speakers. The second point of note is that there have not been many, if any, attempts to automatically detect nasalized vowels, even though researchers have noted that this capability would be very beneficial to help verify the presence of a nasal consonant.

1.3 Summary and Outline of Research

There is clear evidence that the acoustic signal of speech is rich in acoustic information. This implies, that by incorporating more knowledge about the acoustic characteristics of speech sounds, automatic phonetic recognition performances have the potential to be substantially better than are presently obtained in practice.

A survey of previous acoustic studies reported in the literature indicates however, that while the results clearly establish relevant acoustic properties of the speech sound, they are not always directly applicable to speech recognition systems, due to the manner in which measurements were calculated, or due to the nature of the analysis database itself.

The primary objective of this thesis research is the characterization and quantification of nasal consonants and nasalized vowels in American English. The secondary objective is to design automatic nasal consonant and nasalized vowel detection systems which incorporate robust acoustic measures of nasality, and operate in a speaker-independent, continuous-speech environment.

The research in this thesis is organized into two stages. First, an acoustic study of nasal consonants and nasalized vowels is conducted. The main goal of this study is to observe, and quantify the observations made by previous studies, using a large database of natural utterances. Chapter two describes the methodology used for the acoustic study, and chapter three presents the results of the data analysis.

The second stage of this research is concerned with the automatic detection of nasal consonants and nasalized vowels in continuous speech. Chapter four describes and evaluates the detection system, and reports on a set of experiments designed to determine the perceptual merit of the system decisions.

Chapter five presents a summary of the thesis. In addition, suggestions for further research are discussed.

Chapter 2

Data Analysis Methodology

The acoustic analysis of nasal consonants and nasalized vowels, is performed through a series of experiments, and is conducted on a database of utterances. The design of the database requires that several important issues be considered. These issues are discussed in the next section along with a description of the database construction. The following section describes the data analysis procedures used in the acoustic study, and the final section briefly describes the data analysis facility used for all of the acoustic experiments.

2.1 Database Description

Due to the variability of the speech waveform, any attempt to quantify acoustic characteristics of speech sounds requires a carefully designed database. In the past, the majority of researchers have opted to study speech sounds in restricted environments, such as stressed consonant-vowel sequences embedded in nonsense syllables. The theory behind this methodology is that stressed syllables are probably articulated with greater care and effort, and thereby produce a robust acoustic signal whose features may be extracted more reliably [70], [71].

A study using naturally spoken words however, provides greater insight into the acoustic characteristics of sounds in fluent speech. Also, any quantified observations are more useful for automatic continuous speech recognition, since they give a better indication of the variability of these acoustic characteristics. For these reasons, the database was constructed from real words spliced out of continuous speech.

Once the decision was made to construct the database from naturally spoken words, it became necessary to decide which words to include in the corpus. Since the size of the corpus should be as compact as possible, it was important to create one that was well balanced. Thus, the corpus was created using the following criteria:

- The corpus should contain a diverse sampling of the many possible syllabic and phonetic contexts of nasal consonants in American English For example, the corpus should contain nasal consonants in intervocalic, post-fricative, and homorganic nasal stop consonant environments as found in the words *conic*, *smack*, and *pink* respectively.
- The corpus should contain minimal pairs (tokens which have only one phonetic difference), in order to distinguish which acoustic characteristics belong to the nasal consonant class, and which ones do not. For example, the corpus should contain minimal pairs which differ only by the absence of a nasal consonant, as is found in the words *bent* and *bet*. The corpus should also contain minimal pairs which differ only by the substitution of a similar speech sound for the nasal consonant (such as a glide or a voice bar), as is found in the words *made*, and *bade* or *wade*.
- The corpus should contain minimal pairs which can be used to detect acoustic differences within the nasal consonant class itself, such as in the words *simmer*, *sinner* and *singer*.
- The corpus should contain minimal pairs which can be used to establish acoustic differences between nasal consonants in a poly-syllabic versus mono-syllabic environment. Consider for example, the words m*eat* and *voltmeter*, where the syllable-initial nasal consonant has gone from a primary to a secondary stress position.

The contents of the over 200 word corpus may be found in Appendix A.

To produce a database containing utterances which are truly "naturally spoken", the corpus words should be embedded in sentences with acceptable semantic and syntactic structures. However, this type of recording procedure, besides creating a requirement for a large number of carrier sentences, raises the issue of the effect of local syntax and semantics on the individual words. Accounting for this variability would be difficult with different carrier sentences. For this reason, a common carrier sentence was used for all words, so that the corpus words would always be in the same context in the sentence. In this research, the carrier phrase "*She said* ___ *happily*" was used, since it minimized the amount of coarticulation with any word-initial, or word-final nasal consonant, since the phoneme /h/ is neutral, and the phoneme /d/ cannot form a consonant cluster with word-initial nasal consonants in English.

Recordings were made in a sound-isolated room using a Sony omni-directional, electret microphone (model ECM-50PS), a Shure microphone mixer (model M68FC), and a Nakamichi LX-5 tape recorder. The overall signal-to-noise ratio was approximately 30 dB. Original utterances were stored on cassette tape (TDK SA-C60). For recording purposes, the corpus words were randomized into groups of ten. During the recording sessions, speakers were instructed to read naturally and to take a breath at the ends (as opposed to the middle) of phrases. Speakers were allowed to pause for as long as they wanted between each group of ten, but were asked to read each group continuously. Any mispronounced words were repeated immediately following a group of ten. After the recording session, the first and last utterances from each group of ten were deleted in an attempt to minimize artifacts which can occur at the beginning and end of paragraphs.

The analysis database was made from six native speakers of American English (three male and three female) between the ages of twenty and forty. All 1200 utterances were digitized at 16 kHz/s (16 bit words), and their phonetic transcriptions were manually time aligned with the waveform. The time alignment procedures used for the transcription process are described in detail in Appendix B.

2.2 Data Analysis Procedures

The data analysis of nasal consonants and nasalized vowels are divided into three separate studies of duration, energy, and spectral properties. The following sections elaborate on the types of measurements made in each area, and explain how the calculations are computed.

2.2.1 Analysis of Duration

The primary focus of the durational study is to quantify the effect of phonetic context on the duration of the nasal consonant. Nasal consonant durations have been studied more than any other acoustic characteristic of the nasal consonant. Thus, it is easy to compare the results of previous work to those found in this study. Many studies in the past have restricted themselves to one particular phonetic context, such as homorganic nasal stop consonant clusters. An important contribution made by this work therefore, is to allow a comparison of nasal consonant durations in many different environments.

The duration of nasalized vowels are quantified in order to observe their durations relative to oral vowels. Once again, a comparison will be made with previously reported results.

Calculation of Duration

Duration is relatively simple to compute, since it is defined by the time alignment of the phonetic transcription. Although in the past, it has not always been an easy matter to find the exact boundaries of any given phoneme [44], the use of spectrograms simplifies this task. For instance, the temporal boundaries of the nasal coosonant are relatively easy to establish, since they are usually denoted by sharp spectral changes which occur at the beginning and end of the period of oral closure. In general, boundaries produced by different transcription experts are within 10 msec of each other [35].

2.2.2 Analysis of Energy

There are two procedures used to analyze the energy characteristics of nasal consonants. Since nasal consonants occur next to a vowel in English, the first procedure measures the relative difference in average energy between the nasal consonant and an adjacent vowel. When the nasal consonant occurs in a medial context, the largest energy difference is computed.

From a speech recognition perspective, it would be valuable to know how the distribution of this energy difference of nasal consonants compares with other sounds. This would establish if the energy difference measure has any potential for use in a discrimination task. Thus, comparisons are made to sounds with similar acoustic characteristics to nasal consonants, such as semivowels and voice bars. Figure 2.1 contains spectrograms of the words *hammock*, *cob*, and *lip*. Note that the semivowel /l/, and the voice bar in /b/, have similar acoustic characteristics with the /m/.[1]

[1] Since voice bars are not always immediately adjacent to a vowel, a slightly different procedure was also used to quantify energy. In this case, the energy value in the token is relative to the largest energy in the utterance instead of an adjacent vowel. Both of these procedures were found to produce similar results.

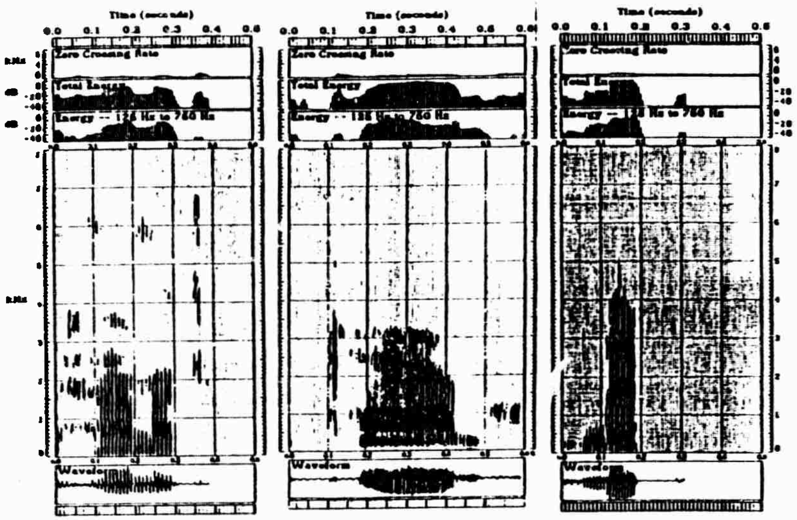


Figure 2.1: Spectrograms of the words *hammock*, *cab*, and *lip*

As nasal consonants are commonly believed to be stable, since their vocal apparatus is held fixed during production, it is of interest to measure how much the energy parameters actually change during the nasal murmur. The stability is measured by computing parameters such as standard deviations of energy values, and average values of first differences of energies in the nasal murmur. Once again, comparisons are made between nasal consonants and similar speech sounds.

Calculation of Energy

Energies are calculated using short-time processing techniques commonly used in digital speech processing. The underlying assumption for the use of these procedures, is that the vocal mechanism is quasi-stationary, in that its acoustic characteristics change slowly with time. Thus, short segments of the speech signal may be isolated and processed as if they were short segments from a sustained

sound. In general, the short-time energy is defined as

$$E_n = \sum_{m=-\infty}^{\infty} (x[m]w[n-m])^2 \qquad (2.1)$$

where $x[n]$ is the speech waveform, and $w[n]$ is a windowing filter, the shape of which can drastically affect the short-time energy function E_n. In general, it is desirable to have a window with an impulse response short enough so that the energy function is responsive to rapid changes in the speech signal. However, the impulse response should also be long enough to provide sufficient averaging of the speech waveform to produce a smooth energy function. Further discussions on windowing may be found in speech processing textbooks [56]. For many digital speech processing applications, a hamming window is used [61]. In this research, a hammirg window of 25 msec duration was used in all of the energy calculations.

Energy in a particular frequency band is computed by taking the dot product of the short-time spectra, $X(e^{j\omega})$, with a frequency window, $Z(e^{j\omega})$, typically of trapezoidal shape (Appendix C contains a discussion of short-time Fourier spectra). Using Parsevals relation for conservation of energy, it can be shown that this procedure is equivalent to producing the short-time energy via equation 2.1 when $x[n]$ is first filtered by a function with frequency response $Z(e^{j\omega})$.

During data analysis, all energies were converted to dB to reduce the sensitivity of the energy function to small changes when the energy signal is large.

For the statistical analysis of energy stability, the average energy of a token, $\bar{E}$, and its standard deviation, σ, computed between two time points n_1 and n_2, are defined as

$$\bar{E} = \frac{1}{n_2 - n_1} \sum_{n=n_1}^{n_2} E_n \qquad (2.2)$$

$$\sigma = \sqrt{\frac{1}{n_2 - n_1} \sum_{n=n_1}^{n_2} (E_n - \bar{E})^2} \qquad (2.3)$$

2.2.3 Analysis of Spectra

Perhaps the most interesting aspect of the acoustic study is the study of the spectral characteristics of the nasal consonants and the nasalized vowels. The spectral analysis performed in this research is carried out in two steps. In the first stage of analysis, the goal is to establish prototypical spectral shapes. From these spectral shapes, it is possible to hypothesize general spectral characteristics of the nasal consonant or nasalized vowel.

The next step in the analysis is to develop algorithms which are able to automatically extract the properties observed in the prototypical spectral shapes. Due to the variability of the speech signal across speaker and context, the emphasis at this stage is on creating measurements which extract information about *robust* characteristics of the nasal consonant or nasalized vowel. Algorithms which try to measure subtle properties of nasality are often fragile, and sensitive to speaker variability, and hence are avoided wherever possible.[3]

Once measurement algorithms are created, the characteristics of utterances in the database are quantified. As was the case for duration and energy, comparisons will be made between the distributions of nasal consonants and those of similar sounds, and between nasalized and non-nasalized vowels.

Finally, part of the analysis is concerned with measuring the spectral stability of the nasal consonant. While there is clearly a significant spectral change at the transition between a nasal consonant and an adjacent vowel, it is worthwhile to quantify the spectral stability of the nasal murmur itself, and to compare this stability to that of similar speech sounds.

Calculation of Spectra

All spectral analysis is based on smoothed spectra computed with the discrete Fourier transform (DFT). The spectra were computed every 5 msec, and were smoothed by windowing the cepstra with a low-pass window that is constant for the first 1.5 msec, and cosine tapered for the next 1.5 msec. Figure 2.2 illustrates an unsmoothed and a smoothed version of a DFT, taken from a nasalized /i/ in the word *technique*. A discussion on the issues involved in spectral analysis may be found in Appendix C.

Wherever it is desired to compute parameters based on the smoothed spectra itself, the spectra are sectioned into peaks, valleys, and transition regions through the use of the second derivative of the smoothed spectra. Boundaries are located at zero crossings of the second derivative spectral slice. Figure 2.3 illustrates an example of a spectral slice which has been schematized in this manner. From this point it is easy to establish spectral peaks and valleys.

Although there are no formal procedures, there are several methods which can be used to measure spectral change in the speech signal. Ultimately, each technique attempts to measure some difference in consecutive short-time spectra. One simple method consists of observing changes in the first few cepstral coefficients, since by their definition,

$$c[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|X(e^{j\omega})|e^{j\omega n}d\omega \qquad (2.4)$$

they just weight the log spectrum by different shaped cosine windows.

Naturally, there is no reason why the windows cannot be an arbitrarily shaped function. In fact, it is quite often advantageous to shape a window function, $W(e^{j\omega})$, so that it is sensitive to spectral changes in a particular frequency region. One way of computing the spectral change parameter, S, is by taking the normalized dot product of the spectral slice, $X(e^{j\omega})$, with the weighting window, $W(e^{j\omega})$,

$$S = \frac{\vec{X} \bullet \vec{W}}{|\vec{X}||\vec{W}|} \qquad (2.5)$$

[3] The actual algorithms used in the data analysis are described in detail in the following chapter

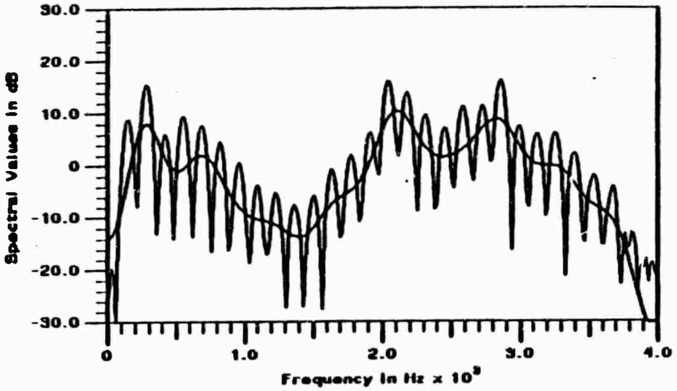


Figure 2.2: An Unsmoothed and Smoothed DFT Spectral Slice

The smoothed DFT spectral slice is computed by windowing the cepstrum with a window that is flat for the first 1.5 msec, and cosine tapered for the next 1.5 msec. This particular example was taken from the /i/ in the word *technique*

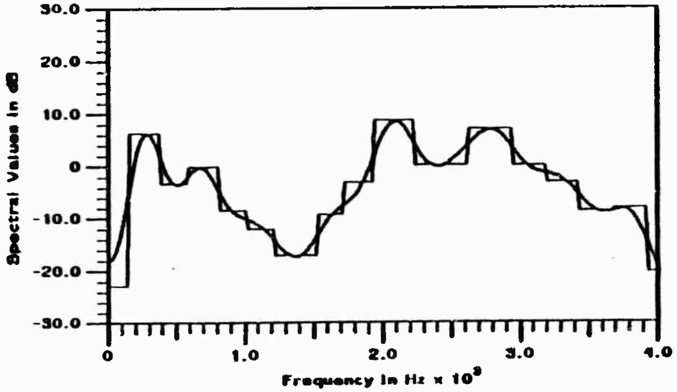


Figure 2.3: A Schematized Spectral Slice

The smoothed DFT spectral slice is schematized into peaks, valleys, and transition regions. Boundaries are located at zero crossings of the second derivative of the smoothed spectral slice.

Note that both of the parameters are treated as vectors in equation 2.5. The magnitude of the spectral change may be computed by taking a first difference of this function.

Spectral Averaging

As previously mentioned, the first stage of the spectral analysis of nasal consonant and nasalized vowel spectra establishes prototypical spectral shapes. Since nasal consonants bave little flexibility in the manner in wbich they are produced, one might expect that for a given speaker, and a given place of articulation, the murmur spectra could be averaged togetber witbout a significant loss of information. This argument can be extended to steady state vowels as well. The validity of this procedure is indicated by the size of tbe variance in tbe spectral average. There are otber, more sophisticated forms of clustering or data reduction such as k-nearest-neighbor, or principal component analysis, wbicb bave been used successfully in tbe past for speech sounds [30], [60], [72]. However, since tbe objective of this first stage is mainly for qualitative observation, and not for quantification, a more sopbisticated analysis procedure is not pursued.

For analysis, spectra are pre-empbasized, and computed from a windowed cepstrum. As well, the spectra are all normalized with respect to total energy so that individual energy offsets are eliminated. Analysis is restricted to one speaker at a time, in order to eliminate speaker variability.

For nasal consonant analysis, statistics are gathered by collecting multiple spectra from all of tbe nasal murmurs. Figure 2.4 shows multiple spectra for /m/ for a female speaker. Note tbat there appears to be common characteristics among the many spectra, suggesting that averaging is reasonable in these circumstances. In pilot studies, there were actually two different averaging procedures which were evaluated. Figure 2.5 shows the average spectra obtained from collecting multiple spectra from each nasal murmur, while figure 2.5 shows the average spectra obtained by collecting a single average spectra from each nasal murmur. In the figures, the thick line is the mean spectral shape, and the outer two lines are one standard deviation away. As can be seen, the average spectral shapes are very similar. The standard deviation of the multiple spectra averaging technique is slightly larger. This is to be expected however, since there are a larger number of spectra included in the averaging. The fact that the two averaging techniques yield similar results illustrates that the spectral characteristics of the nasal murmur are quite stable against time, especially at low frequencies. Since both spectral averaging techniques yielded similar results, the multiple spectral averaging procedure was used for all data analysis since it gave a better indication of the variance of the spectral shapes.

The same multiple spectra averaging technique is used for analysis of nasalized vowels. Even though the averaging procedure is quite informative, care must be taken in interpreting the average spectra, since nasalization is not a static spectral characteristic, but often changes the duration of a vowel. Figure 2.6 illustrates the case for the word *mitt*, where the spectral characteristics of the low resonance region on the left side of the vowel, are clearly different from those on the right.

2.3 Data Measurement Facility

Data analysis is performed with the Spire and SpireX facilities available on MIT Lisp machine workstations [69]. SpireX is a statistical analysis package which allows the user to perform acoustic-phonetic experiments on a large body of utterances. Using SpireX, a typical experiment proceeds in five steps, each of which is described in the following paragraphs.

Catalogs

A user first specifies a *catalog* of utterances to be used for the experiment. A

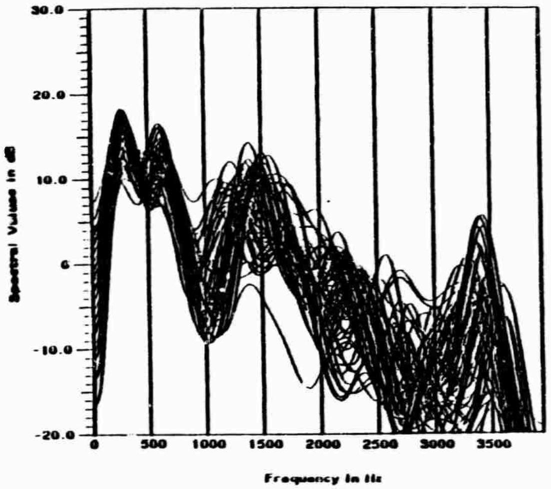


Figure 2.4: Multiple Spectra of an Intervocalic /m/

This display presents an overlay of the normalized smoothed spectra occurring during the nasal murmur of an intervocalic /m/ for a female speaker.

context of interest, SpireX searches through the catalog for instances of the desired phonetic context. Each such instance is known as a *sample*. The phonetic context is specified as a sequence of named regions, each of which consists of a given phonetic pattern. For example, a region could specify a class of phonemes, a specific phoneme, or a more complicated pattern. Thus, to collect a sample set of all nasal stop consonant clusters in a catalog, where the nasal and stop are homorganic and dental, the search specification could consist of a sequence of the three regions *vowel*, *nasal*, *stop*, where *vowel* is any vowel, *nasal* is an /n/, and *stop* is a /d/, or a /t/.

Once the search is completed, each region is associated with a time-interval for each sample in the sample set The region names are used as arguments in later steps of the experiment to reference these time-intervals.

Computations

After the search is complete, the user then specifies a set of computations to be performed on each sample. Computations are usually supplied with search regions and Spire attribute names as arguments. A computation then performs statistical measurements of the attributes in the time-intervals specified by these regions. Typical computations include averages, maximums, and durations. Although computations are usually specified in terms a menu driven interface, users are also allowed to define their own computations, although this requires some knowledge of SpireX. In the nasal stop consonant cluster example, typical computations might include the durations of the *nasal*, and *vowel* regions, and a binary computation which indicates if the *stop* is voiced, or voiceless.

Filters

Filters are logical computations which are used to separate the sample set into groups Only samples which match a filtering specification are included in the

statistical analysis. Thus, for the nasal stop example, the voicing computation could be used to filter the sample set. This would allow the user to separate the statistics of the *nasal*, and *vowel* duration computations of voiced stops from those of voiceless stops.

Display

Once the sample set has been filtered, it is possible to perform a statistical analysis on specified computations and view or tabulate the results. The display capabilities include histograms, scatter plots, and statistical summaries.

2.4 Chapter Summary

This chapter discussed the methodology used for the acoustic analysis of nasal consonants and nasalized vowels. The major points of the chapter were,

1. Data analysis is accomplished by performing a series of experiments on a database of utterances.

2. The database consists of over 1200 words, excised from continuous speech, and recorded from six speakers, three male, and three female. All of the recorded words were digitized, and their phonetic transcriptions were aligned with the speech waveform.

3. Data analysis is divided into a study of nasal consonants and nasalized vowels. In each study, measurements are made of the duration, energy, and spectral characteristics.

4. Data analysis is performed using the SpireX statistical analysis facility.

Chapter 3

Data Analysis

This chapter presents the results of the data analysis experiments carried out on the utterances in the database. The analysis was performed separately on the nasal consonants, and nasalized vowels. The results of each are presented in the following sections.

3.1 Analysis of Nasal Consonants

3.1.1 A Study of Nasal Consonant Duration

Minimal Pair Experiments

As a first step at analyzing the effects of phonetic context on the duration of the nasal consonant, the differences of minimal word pairs, such as *bend/bent*, or *mack/smack*, were observed. The minimal pairs were restricted to monosyllabic words in order to eliminate possible secondary effects introduced in multi-syllable environments (the nasal murmur in the word *picnic* is much shorter than in the word *nick* for instance). The results of these minimal pair experiments, which included all of the speakers in the database, are presented in figure 3.1, and are summarized below:

1. The durations of word final nasals are lengthened when clustered with a voiced stop consonant (VS), such as for the minimal pair *ben/bend*. For the ntterances in the database, the average duration increase was 10 msec, or 20% of the duration of the singleton nasal. Also evident from the figure is that word final nasals are shortened when clustered with an unvoiced stop cnnsonant (US), such as for the minimal pair *ben/bent*. The average duration decrease was found to be 20 msec, or 40% of the singleton nasal duration.

2. The same trends are observed when word final nasals are clustered with a fricative. When the fricative is voiced (VF), as in the minimal pair *one/ones*, the average duration increase is 28 msec, or 35% of the singleton nasal duration. When the fricative is unvoiced (UF), as in the minimal pair *one/once*, the average duration decrease is 18 msec, or 30% of the singleton nasal duration.

3. The duration of word initial nasals are shortened when clustered with an unvoiced fricative consonant (F), such as for the minimal pair *nack/snack*. The average duration decrease was observed to be 40 msec, or 50% of the singleton nasal duration. Since there are no word initial voiced fricative nasal clusters in American English, the opposite trend could not be observed.

4. As implied by the previous experiments, the duration of a nasal in a word final consonant cluster is longer when the clustering consonant is voiced, than when it is voiceless. When the difference for stop consonants (VUS), such as for the minimal pair words *canned/can't*, was observed, the average difference in duration of the nasal consonant was 25 msec. Note that only the phonemes /t/ and /d/ were relevant here, since there are no word final nasal stop consonant clusters with the phonemes /b/, and /g/.

5. The same trends were observed in word final nasal fricative clusters (VUF), such as for the minimal pair words *ones/once*. The average difference was measured to be 40 msec.

Two interesting observations were made from these experiments. Using the knowledge that voiced stop consonants have shorter stop gaps than voiceless stop consonants [76], a simple guideline was established for distinguishing voicing in nasal stop consonant clusters, as shown in figure 3.2. It was found that when the nasal murmur occupied over 80% of the duration of the nasal murmur and stop gap, the stop consonant was voiced 90% of the time. If the fraction was less than 0.7 however, the stop consonant was unvoiced 87% of the time. This observation included poly-syllabic words as well.

As shown in figure 3.3, this same observation was found to hold true for stop nasal consonant sequences, such as /pm/, in the word *chipmunk*. When the fraction was less than 0.4, the stop consonant was unvoiced 83% of the time. When the fraction was greater than 0.5 the stop consonant is voiced 88% of the time.

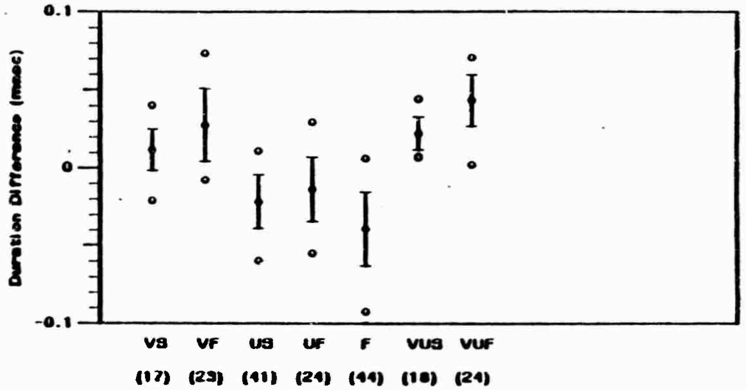


Figure 3.1: Statistical Summary of Minimal Pair Experiments

This display summarizes the results of minimal pair experiments which measured differences in the duration of nasal consonants in two different contexts. From left to right, the contexts are: word final vs. nasal voiced stop (VS), word final vs. uasal voiced fricative (VF), word final vs. nasal unvoiced stop (US), word final vs. nasal unvoiced fricative (UF), word initial vs. unvoiced fricative nasal (F), nasal voiced stop vs. nasal unvoiced stop (VUS), and nasal voiced fricative vs. nasal unvoiced fricative (VUF). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

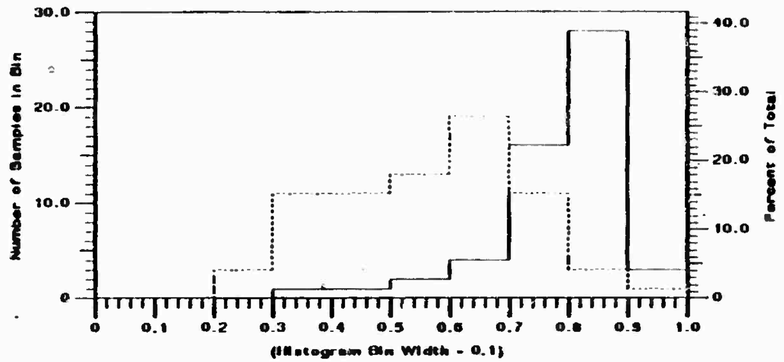


Figure 3.2: Voiciog Discrimination in Nasal Stop Consonant Sequences

The solid lines outlioe the fraction of time that tbe nasal murmur occupies the total oasal oiurmur and stop gap duratioo of voiced stops (55 tokens). The dashed lines outline the same fraction for voiceless stops (72 tokena).

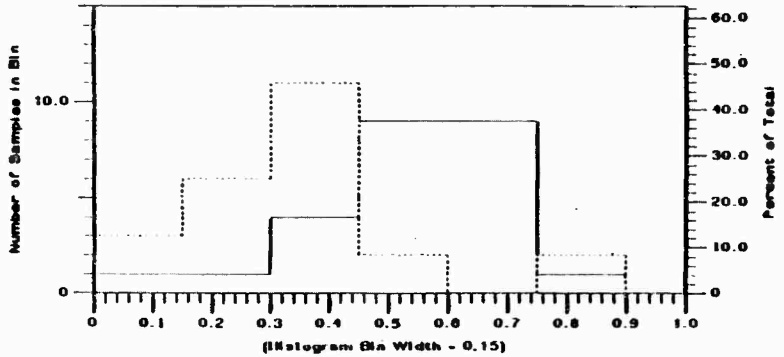


Figure 3.3: Voicing Discrimination in Stop Nasal Consonant Sequences

The solid lines outline the fraction of time that the nasal murmur occupies the total stop gap and nasal murmur duration of voiced stops (25 tokens). The dashed lines outline the same fraction for voiceless stops (24 tokens).

General Results

In an attempt to establish global duration values of different contexts, the database was reduced to a set of monosyllabic words, with the exception of intervocalic nasals These words were subdivided into broad contexts. It was found that the duration of nasal murmurs produced by male speakers were affected by context slightly more than females. Global duration valnes for the two groups are shown in figures 3.4 and 3.5. Note that the average durations of female speakers are greater than the male counterparts in every context. Figure 3.6 summarizes the durations of nasal murmurs in a singleton environment, and those in a cluster with another consonant, for all speakers in the database. Nasal murmurs in a singleton environment had an average duration of 65 msec. Nasal murmurs in a cluster with a voiceless consonant had an average duration of 40 msec, while those in a cluster with a voiced consonant had an average duration of 75 msec. Thus, voiceless clusters tend to shorten nasal murmnr duration, while voiced clusters tend to lengthen nasal murmur duration.

In general, fricative nasal consonant clusters had the shortest nasal murmur durations in the database. Figure 3.7 illustrates the distributions of the nasal murmur duration of prevocalic nasals in a syllable initial position, and in a fricative cluster. Fricative consonant clusters also exhibited a period of epenthetic silence between the fricative and the nasal murmur, which was due to a mistiming of the articulators. This period of silence is a very robust acoustic cue for detecting the presence of a nasal consonant (it can also be present in fricative glide clusters such as *slack*) when the nasal murmur is very short. The average duration of the period of silence was found to be 30 msec, as indicated in figure 3.8.

Discussion

Previous studies of nasal murmur durations have been primarily concerned with homorganic nasal stop consonant clusters [79], [62]. All of these investigations have shown that the duratioo of the nasal murmur is substantially longer when preceding a voiced stop than a voiceless s.op. Minimal pair differences range from 25 to 70 msec. The average values fouud in this research are in the lower end of these values. This is probably because many of these studies measured the durations of nasal consonants in stressed, monosyllabic words, sometimes spoken in isolation. Thus, one would expect tokens spliced out of continuous speech to have shorter durations.

Other researchers have noted the differences in duration of the nasal murmur between male and female speakers [77]. The general finding is that when nasals form a cluster with another consonant, the nasal murmur duration of female speakers is not affected to the same degree as those of male speakers.

It should be emphasized that the majority of the duration statistics were gathered on a subset of the database. With the exception of intervocalic nasal consonants, poly-syllabic words were not included. As one might expect, a minimal pair experiment found that the durations of nasal murmurs in a poly-syllable environment, were shorter than those in a mono-syllable environment, as illustrated in figure 3.9. Thus it would be difficult to apply the knowledge of duration of nasal murmurs to the field of speech recognition, unless one was able to obtain details of the particular context of the nasal consonant under consideration. The fact that the rate of speech itself can vary substantially, further limits the usefulness of duration as a speech recognition parameter.

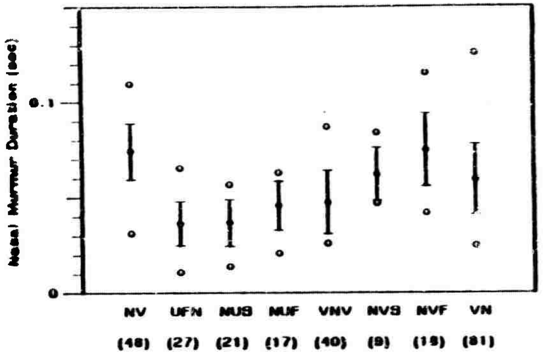


Figure 3.4: A Summary of Nasal Consonant Durations for Male Speakers

This display summarizes nasal murmur durations of male speakers for particular phonetic environments. From left to right they are: singleton prevocalic nasals (NV), fricative nasal clusters (UFN), nasal unvoiced-stop consonant clusters (NUS), nasal unvoiced-fricative clusters (NUF), intervocalic nasals (VNV), nasal voiced-stop consonant clusters (NVS), nasal voiced-fricative clusters (NVF), and singleton post-vocalic nasal consonants (VN). The average value is indicated by a filled circle. Vertical lines indicate one standard deviation. The open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

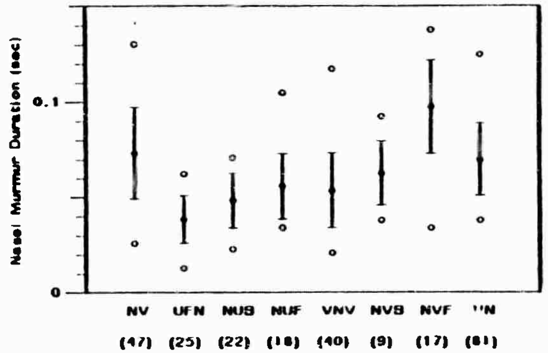


Figure 3.5: A Summary of Nasal Consonant Durations for Female Speakers

This display summarizes nasal murmur durations of female speakers for particular phonetic environments. The contexts are the same as those described in figure 3.4. The average value is indicated by a filled circle. Vertical lines indicate one standard deviation. The open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

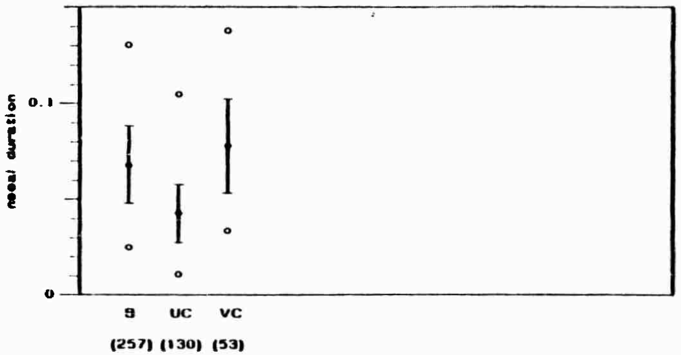


Figure 3.6: A Summary of Nasal Consonant Durations

This display summarizes nasal murmur durations of nasal consonants in a singleton environment (S) as opposed to those in a cluster with a unvoiced consonant (UC) or a voiced consonant (VC). The average value is indicated by a filled circle. Vertical lines indicate one standard deviation. The open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

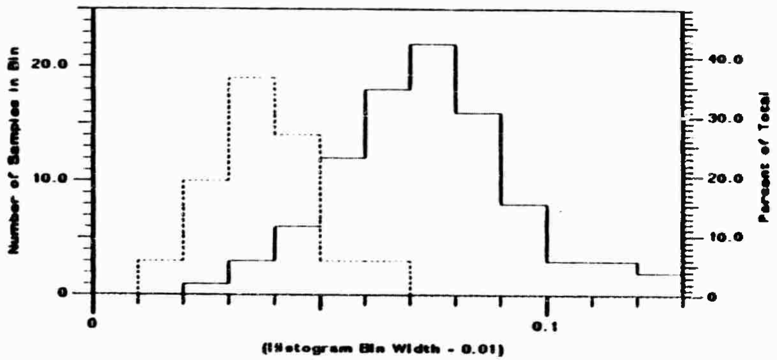


Figure 3.7: Durations of Prevocalic Nasal Consonants

The solid lines outline the duration of nasal consonants in a word initial position (95 tokens). The dashed lines outline the durations of nasal consonants which form a fricative nasal cluster (52 tokens). Thus, this display compares words like *knit*, and *snit*.

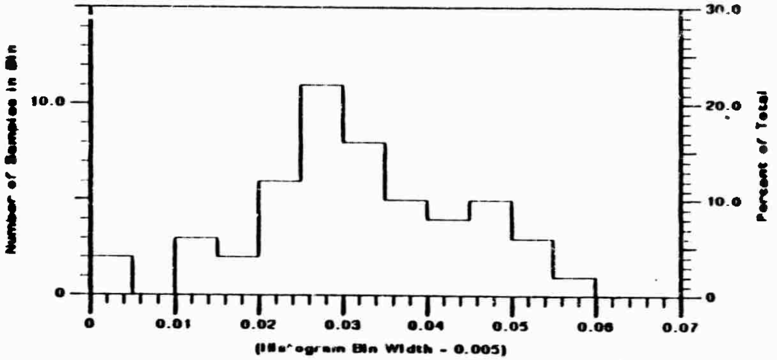


Figure 3.8: Epenthetic Silence Duration of Fricative Nasal Consonant Clusters

The solid lines outline the duration of the period of epenthetic silence between the fricative and the nasal consonant in fricative nasal clusters, as found in the word *snit* (50 tokens).

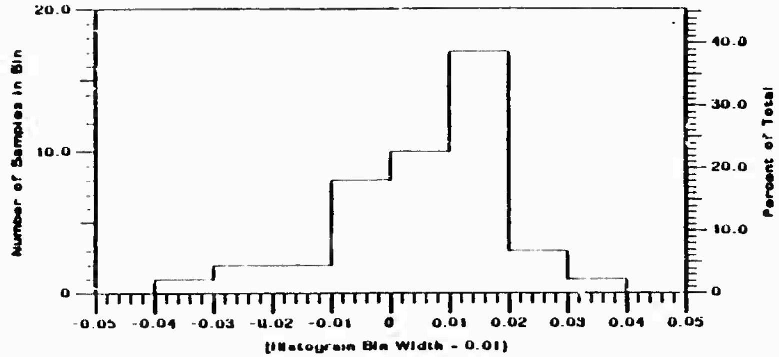


Figure 3.9: Duration Differences of Nasal Consonants between Mono and Poly-Syllabic Words

The solid lines outline minimal pair duration differences between the nasal murmur in a mono-syllabic word and that in a poly-syllabic word, as in the pair *bend/bending* for example (44 tokens).

3.1.2 A Study of Nasal Consonant Energy

The first energy experiment conducted measured an energy difference, calculated by subtracting the average total energy in the nasal murmur from the average total energy in the adjacent sonorant. Figure 3.10 contains a histogram of this energy difference, plotted in dB, for all of the nasal consonants in the database. Since this energy difference is almost always positive, it can be concluded that the nasal murmur is consistently weaker than an adjacent sonorant.

On closer inspection, there are several other observations which may be made about energy differences. Some of these have been illustrated in figure 3.11, which presents a statistical summary of energy differences of nasal consonants in different contexts. As indicated in the figure, there appears to be only minor differences in the energy difference due to vowel quality (front, back, high, or low). The most significant separation appears to be between low back vowels, which have an average energy difference of around 11 dB, and high front vowels, which have an energy difference of 7 dB. This observation is more likely due to the fact that low back vowels have more energy than high front vowels, rather than there being any difference in nasal consonant strengths in these two contexts [12], [23].

As illustrated in figure 3.12, nasal consonants in a medial position between two sonorant regions, have a slightly smaller energy difference, ofaround 6 dB, than nasal consonants in other contexts, typically around 10 dB. This is probably due to the fact that medial nasals have strong energy throughout the murmur, since they are surrounded by two sonorants which have strong energy, and do not taper off as would nasals in other contexts. This observation is reinforced by measurements of the nasal murmur stability, discussed shortly, which indicate that the energy of medial nasals is quite steady.

Figure 3.12 also compares the value of the energy difference of the nasal consonants to similar sounds such as liquids and glides, and voice bars (the common name for the period of closure of voiced stop consonants), which are also adjacent to a sonorant region. Although there is some overlap in the distribntions, it is clear that on average, nasal consonants have a greater drop in energy than the liquids or glides, and have a smaller difference than voice bars.

As was mentionned in chapter 2, it was not possible to measure a relative energy difference for all voice bars, since many were not immediately adjacent to a sonorant region, being separated by a stop consonant release. However, it is possible to compare the average energy of these two groups. As shown in figure 3.13, the average energy of isolated voice bars tends to be mucb weaker than voice bars adjacent to a sonorant, and can be discriminated from most nasal consonants on the basis of energy alone. Figure 3.14 presents a statistical summary of the total energy of nasal consonants and similar sounds.

The next parameter which was observed was the energy stability of tbe nasal consonant. This was measured by calculating the average value of tbe first difference of the energy in the middle 50% of the nasal murmur. This measure is proportional to calculating the standard deviation of the energy in the murmur. Figure 3.15 illustrates a histogram of the average difference for all of the nasal consonants in the database. Figure 3.16 presents a statistical summary of the average difference for similar speech sounds.

Discussion

The most important point of the analysis of nasal consonant energy, is tbat nasal consonants tend to be weaker than adjacent sonorants by an average of 10 dB. This result agrees with previous studies of the nasal consonants [24].

For speech recognition, the energy of the nasal consonant has the potential to be a useful parameter, since nasal consonants tend to be stronger than voice bars, and weaker than semivowels.

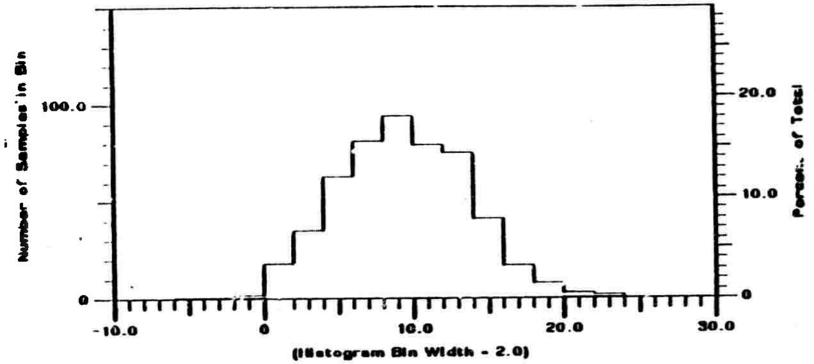


Figure 3.10: Energy Difference of Nasal Consonants

This figure contains a histogram of the energy difference between a nasal consonant, and an adjacent sonorant (520 tokens). Values are plotted in dB.

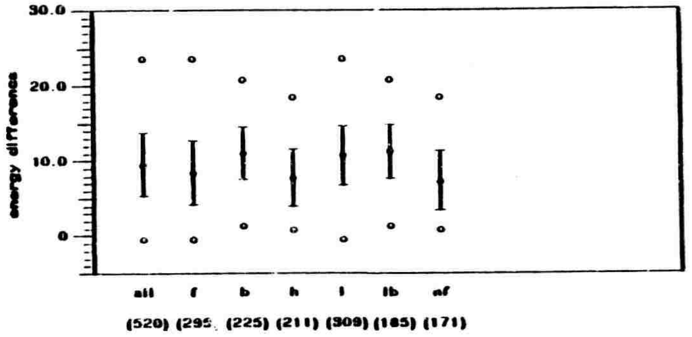


Figure 3.11: Energy Difference Statistics due to Vowel Quality

This display summarizes energy differences of nasal consonants in different contexts. From left to right, they are: all nasal consonants (all), nasals adjacent to front vowels (f), back vowels (b), high vowels (h), low vowels (l), low back vowels (lb), and high front vowels (hf). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display maximum and minimum values. The number of samples in each context are indicated below the display.

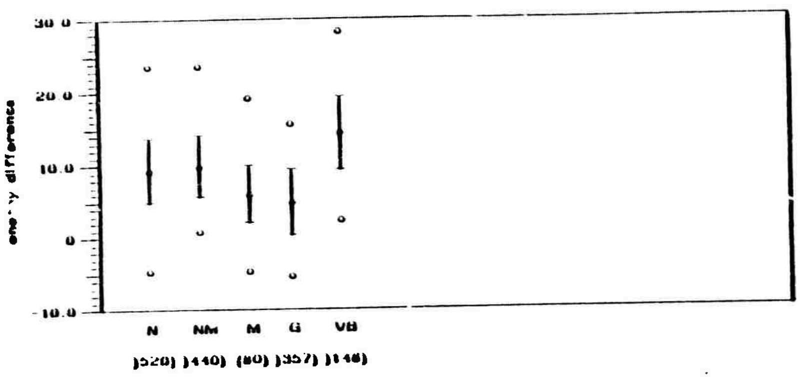


Figure 3.12: Energy Difference Statistics of Similar Sounds

This display summarizes energy differences of nasal consonants and similar sounds. From left to right, they are: all nasal consonants (N), non-medial nasals (NM), medial nasals (M), liquids and glides (G), and voice bars (VB). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation and the open circles display maximum and minimum values. The number of samples in each context are indicated below the display.

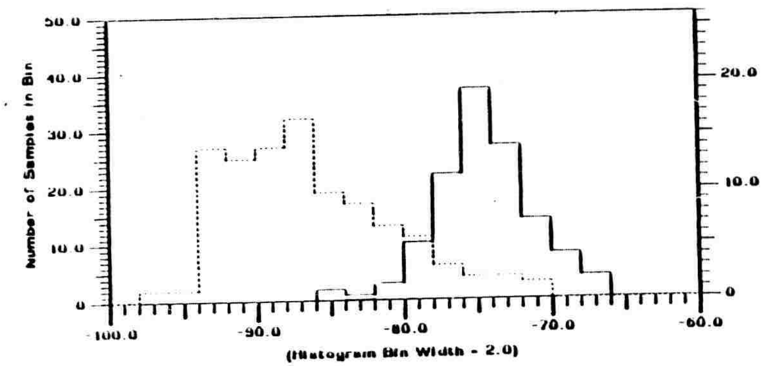


Figure 3.13: Energy of Voice Bars

The solid lines outline average energy for voice bars adjacent to a sonorant (146 tokens). The dashed lines outline the average energy of voice bars not adjacent to a sonorant (192 tokens).

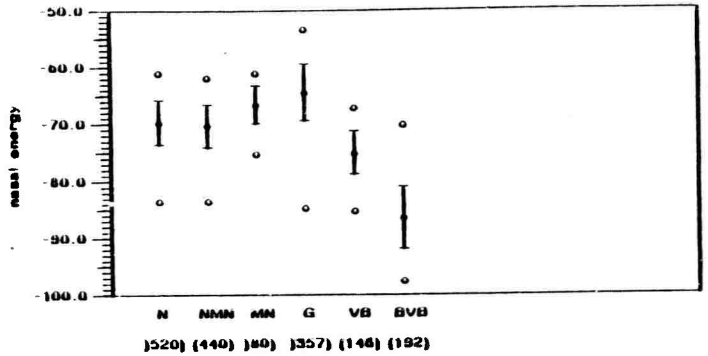


Figure 3.14: Statistics of Energy

This display summarizes energy differences of nasal consonants and similar sounds. From left to right, they are: all nasal consonants (N), non-medial nasals (NMN), medial nasals (MN), liquids and glides (G), voice bars adjacent to a sonorant (VB), and voice bars not adjacent to a sonorant (BVB). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

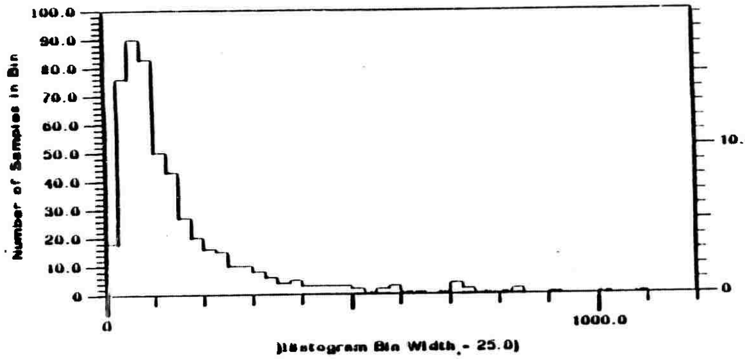


Figure 3.15: Energy Stability of Nasal Consonants

This figure contains a histogram of the average first difference of energy, calculated in the middle region of the nasal murmur (520 samples). Values are plotted in dB per second.

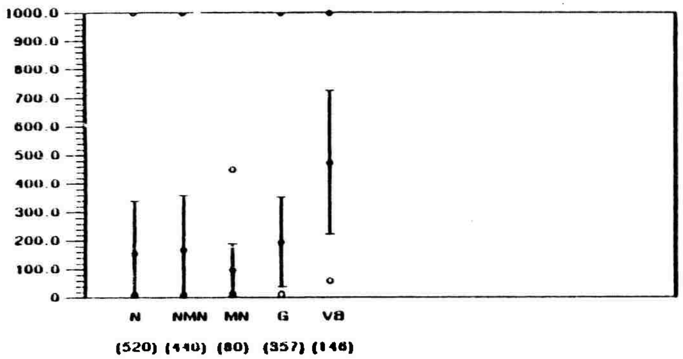


Figure 3.16: Statistics of Energy Stability

This display summarizes the average energy change of nasal consonants and similar sounds. From left to right, they are: all nasal consonants (N), non-medial nasals (NMN), medial nasals (MN), liquids and glides (G), and voice bars (VB). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

3.1.3 A Study of Nasal Consonant Spectra

Once the spectral shape of the nasal murmurs were normalized with respect to total energy, it was possible to measure an average spectral shape using the techniques described in chapter 2. Figures 3.17 and 3.18 show average spectra of the three nasal consonants for one speaker. In general, the spectral shapes of the nasal consonants were found to be highly speaker dependent. This is not surprising, since the size of the nasal and sinus cavities can vary greatly from speaker to speaker. Although subtle differences could be detected between the three nasal consonants for any given speaker, all three nasal consonants tended to have similar spectral shapes, as indicated by these figures. This observation is in agreement with that made by Fujimura, who also found little differences among the magnitude spectra of the three nasal consonants [15].

In general, nasal consonant spectra were characterized by a low frequency energy which dominated the spectrum. Several measures were made in order to quantify this property.

Figure 3.19 plots the frequency of the largest peak in the spectrum for all of the nasal consonants in the database. As may be seen, the low frequency energy was not only nearly always the largest in the spectrum, it was nearly always centered between 200 and 350 Hz as well. Figure 3.20 displays the results of a measurement which calculated the percentage of the time that a nasal consonant had a resonance centered between 200 and 350 Hz (all values are scaled by 100). The majority of nasal consonants had percentage values close to 1.0. Figure 3.20 also plots this percentage for semivowels as well. Figure 3.21 presents a statistical summary of the percentage values for nasals, semivowels, and voice bars. From these figures, it may be concluded that the presence of a low frequency resonance is a necessary, but not sufficient, condition for the identification of a nasal consonant. In other words, if a token does not have a value near 1.0 for the calculation, it is extremely unlikely that it is a nasal consonant.

As previously mentioned, this low resonance energy dominates the overall spectrum of the nasal murmur. Figure 3.22 displays the results of a measure which calculated the relative amount of energy in the low frequency region of the spectrum (below 500 Hz) for all of the nasal consonants, and semivowels. This measure may also be obtained by plotting the normalized amplitude of the low resonance directly. Clearly, the majority of the energy in the nasal consonant is found in the low frequency region. Figure 3.23 presents a statistical summary of this measure for nasal consonants, semivowels, and voice bars.

The final characteristic of the low resonance which was quantified, was an abrupt decrease in energy in the frequencies immediately above the low resonance. Figure 3.24 displays the results of a measure which calculated the amount of low frequency energy (below 350 Hz) relative to local adjacent energy (350 to 1000 Hz). This measure was not overly sensitive to the actual locations of the frequency boundaries. This measure could also be obtained by spectral weighting functions, such as center of gravity measures in the low frequency region. Figure 3.25 presents a statistical summary of this measure for nasal consonants, semivowels, and voice bars. From this figure, it is apparent that semivowels have less of a drop than nasal consonants, and voice bars have slightly more. In fact, this measure is very effective in seperating nasal consonants from most semivowels.

Finally, a measure of the spectral stability of the nasal consonant spectra was also made. As was indicated by figure 2.5, the spectra of nasal consonants were found to be quite stable at frequencies below 1000 Hz. There are several ways that this can be measured, including measuring the standard deviation from a spectral average, or a spectral weighting function, such as the center of mass. Figure 3.26 displays a histogram of the average deviation of the normalized low frequency energy (below 1000 Hz). The distribution of voice bars is also displayed for comparison. Figure 3.27 presents a statistical summary of this measurement for similar sounds.

Discussion

The analysis of the nasal consonant spectra primarily verified the results of previous studies, which indicated that the spectrum is dominated by a low frequency energy around 300 Hz.

There were several properties of nasal consonants which were difficult to quantify successfully. For instance, it is commonly known that the nasal consonant has several higher frequency resonances, and that the resonance bandwidths are generally higher than in vowel-like sounds. Furthur, nasal consonants have an antiformant, whose frequency location depends on the place of articulation. The problem with attempting to measure any of these parameters is that resonances do not always show up as peaks in the magnitude spectrum, and antiformants will not necessarily show up as valleys in the spectrum. This phenomenon results from pole zero cancellation, as Fujimura illustrated.

For speech recognition purposes, the most robust spectral property of the nasal consonant would appear to be a steady low frequency resonance, which is centered between 200 and 350 Hz. The most useful characteristics of this resonance are the percentage and height measures, since they are able to discriminate nasal consonants from other sounds with similar acoustic properties. The measure of low resonance amplitude is more useful at discriminating between nasal consonants and sounds which do not have a predominance of low frequency energy.

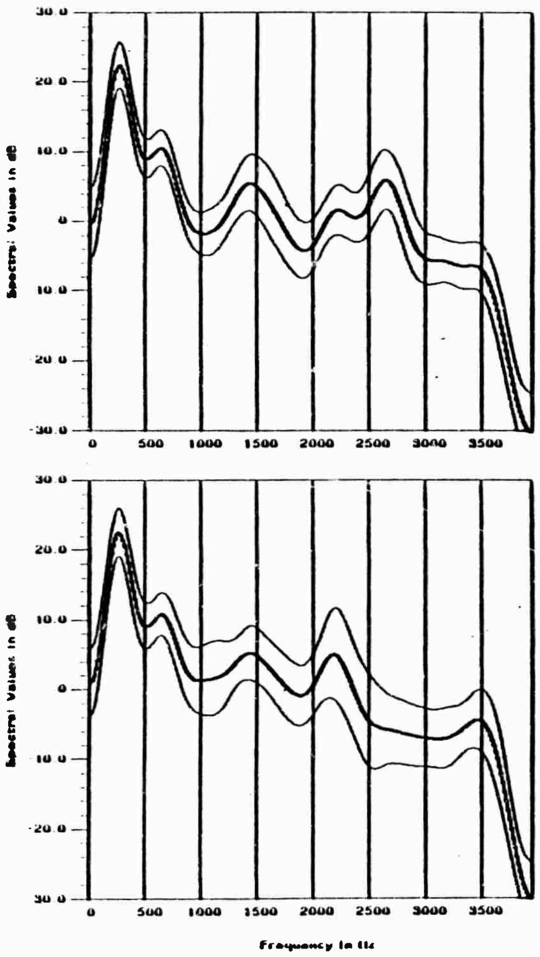


Figure 3.17: Average Spectral Shape of /n/, and /m/

This top display presents a statistical summary of the normalized smoothed spectra of the nasal consonant /n/, for a male speaker. The bottom display presents a summary of an /m/, spoken by the same speaker. The average spectral shape, shown by the dark line, is surrounded by lines which represent one standard deviation from the mean.

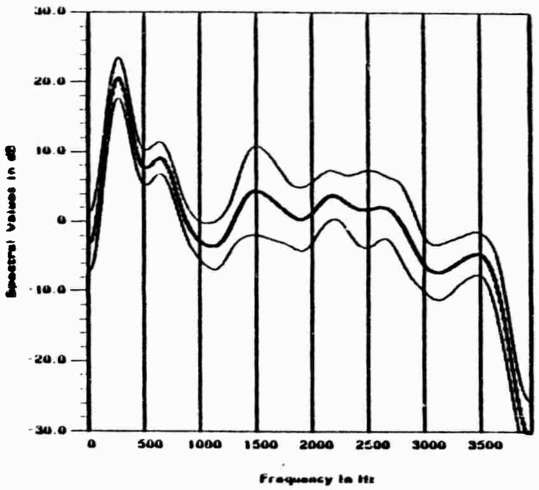


Figure 3.18: Average Spectral Shape of /ŋ/

This display presents a statistical summary of the normalized smoothed spectra of the nasal consonant /ŋ/, for a male speaker. The average spectral shape, shown by the dark line, is surrounded by lines which represent one standard deviation from the mean.

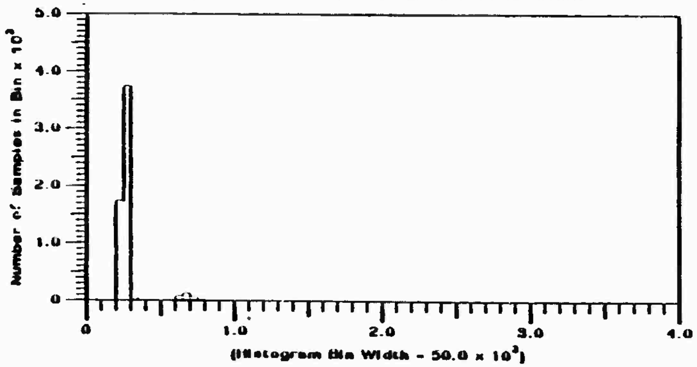


Figure 3.19: Frequency of Largest Spectral Peak in the Nasal Consonant

This display contains a histogram of the frequency of the largest spectral peak in the nasal consonant (6092 tokens). Values were collected for multiple spectra from each nasal consonant, and are plotted in thousands of Hz.

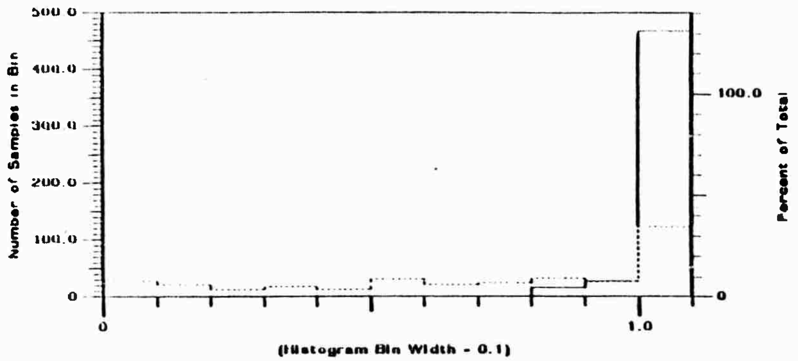


Figure 3.20: Low Resooance Percentage

This display contains a histogram of the percentage of time io a nasal consonant that there was a low frequency resonance centered in between 200 and 350 Hz. The solid lines are the distributions of nasal consonants (520 tokens). The dashed lines are the distributions of semivowels (357 tokens). Values betweeo 1.0 and 1.1 have a value of 1.0

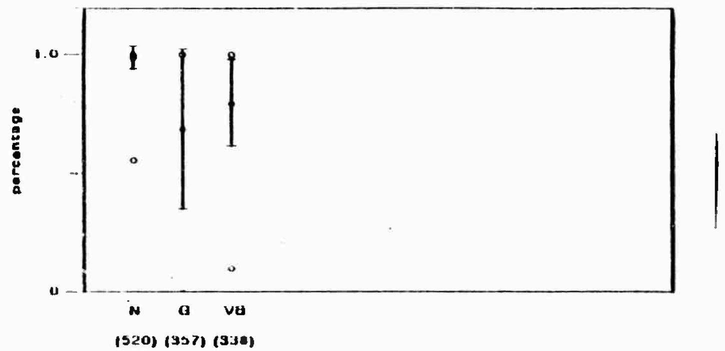


Figure 3.21: Statistics of Low Resnnance Percentage

This display summarizes the low resonance percentage of nasal consonants and similar sounds. From left to right, they are: all nasal consonants (N), liquids and glides (G), and voice bars (VB). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

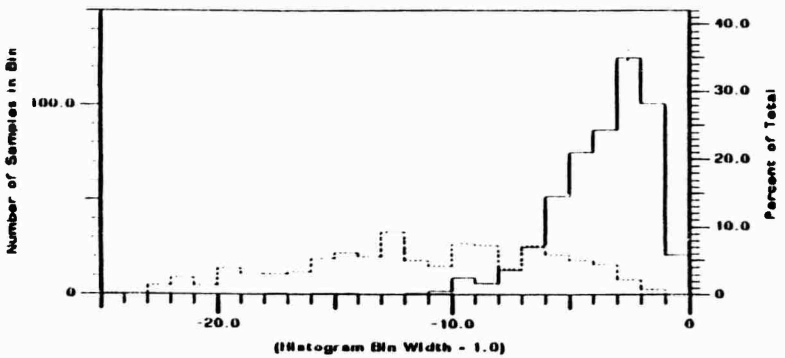


Figure 3.22: Low Resonance Amplitude

This display cootains a histogram of the relative amplitude of low frequency energy. The solid lines are the distributions of the oasal consonants (520 tokens). The dashed lioes are the distributions of semivowels (357 tokens). Values are plotted in dB.

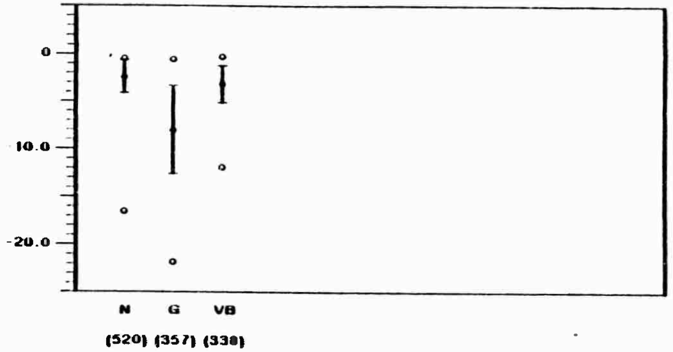


Figure 3.23: Statistics of Low Resonance Amplitude

This display summarizes the lnw resonance amplitude of nasal consonants and similar sounds. From left to right, they are: all oasal coosonants (N), liquids and glides (G), and voice bars (VB). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

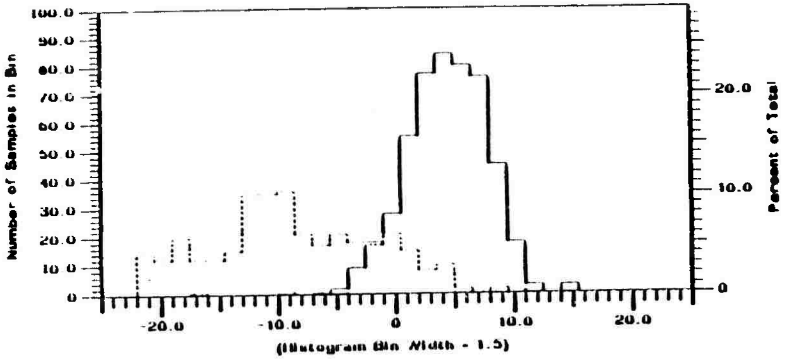


Figure 3.24: Low Resonance Height

This display contains a histogram of the local relative amplitude of the low frequency resonance. The solid lines are the distributions of the nasal consonants (520 tokens). The dashed lines are the distributions of semivowels (357 tokens).

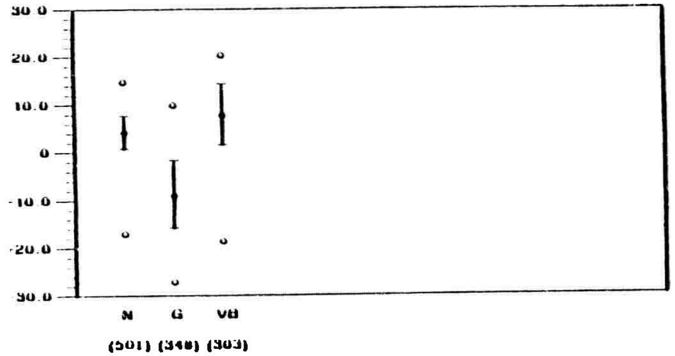


Figure 3.25: Statistics of Low Resonance Height

This display summarizes the low resonance height of nasal consonants and similar sounds. From left to right, they are: all nasal consonants (N), liquids and glides (G), and voice bars (VB). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

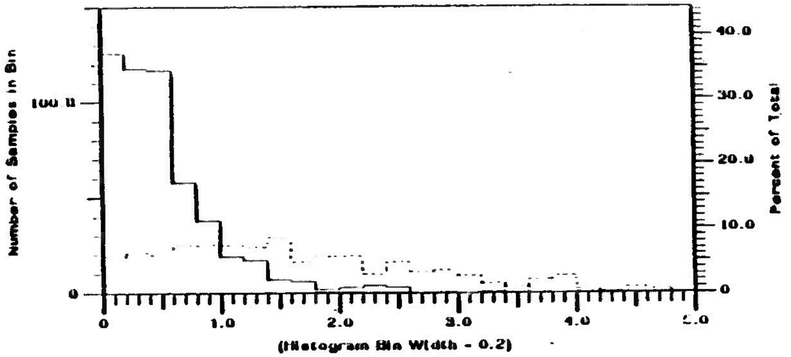


Figure 3.26: Spectral Stability

This display contains a histogram of the standard deviation of relative amplitude of low frequency energy. The solid lines are the distributions of the nasal consonants (520 tokens). The dashed lines are the distributions of semivowels (357 tokens). Values are plotted in dB.

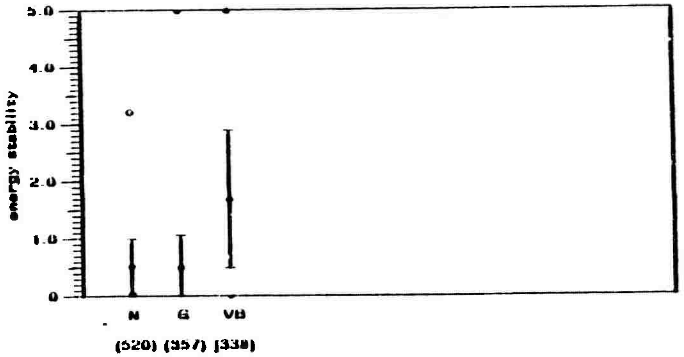


Figure 3.27: Statistics of Spectral Stability

This display summarizes the spectral stability of nasal consonants and similar sounds. From left to right, they are: all nasal consonants (N), liquids and glides (G), and voice bars (VB). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The number of samples in each context are indicated below the display

3.2 Analysis of Nasalized Vowels

A study of nasalized vowels is more complicated than a study of nasal consonants. Since nasalized vowels are not distinguished phonemically from oral vowels in American English, it is perfectly legitimate to nasalize a vowel in any phonetic context. It is therefore not possible to separate nasalized and oral vowels by a phonetic transcription alone. Nasalization could be established by measuring the airflow from the nasal cavities. Vowels with any nasal coupling would then be easily separated from completely oral vowels. Subsequently, an acoustic study of the speech waveforms cuuld establish properties which separate these two groups of vowels.

However, the goal of this research is to establish properties of nasalized vowels which may be used to help detect the presence of a nasal consonant. Thus, it is more useful to classify vowels using the criterion of whether or not they are adjacent to a oasal consonant. The goal is then to establish acoustic differences between these two groups of vowels, making the acoustic study one of *relative* nasalization. The underlying assumption is that when vowels are next to a nasal consonant, they are nasalized *more* than they would be otherwise. In this research then, nasalized vowels are defined as those vowels adjacent to a nasal consonant, while non-nasalized vowels are those vowels that are not adjacent to a nasal consonant.

Although the results of such a study are potentially beneficial to speech recognitiou, there are several complicating factors. First, in American English, a vowel is often nasalized whenever a nasal consonant is present somewhere in the syllable nucleus, even if it is not immediately adjacent to the vowel. For instance, the /l/ in the word *film* will teud to be nasalized. By the definition used in this research, /l/ would be classified as a non-nasalized vuwel. By its context however, it is likely to be nasalized Since the uature of these vowels is somewhat ambiguous, thry were filtered out of the database in order to reduce the amount of noise they might cause in the measurement distributions. This excluded about 200 vowels from the acoustic analysis. Another alternative would have been to classify them as nasalized vowels, since it is likely that they were indeed nasalized.

Although filtering operations will reduce the number of nasalized vowels in a non-nasal context, it will never eliminate all such cases, as is illustrated for the word *back*, shown in figure 3.28. This is because some speakers tend to naturally nasalize all vowels, and also because low vowels are quite often slightly nasalized, independent of context. Clearly, the challenge of the acoustic study is to establish measures which can automatically differentiate between the /æ/ in *back*, and the /æ/ in a word like *mack*, also shown in figure 3.28.

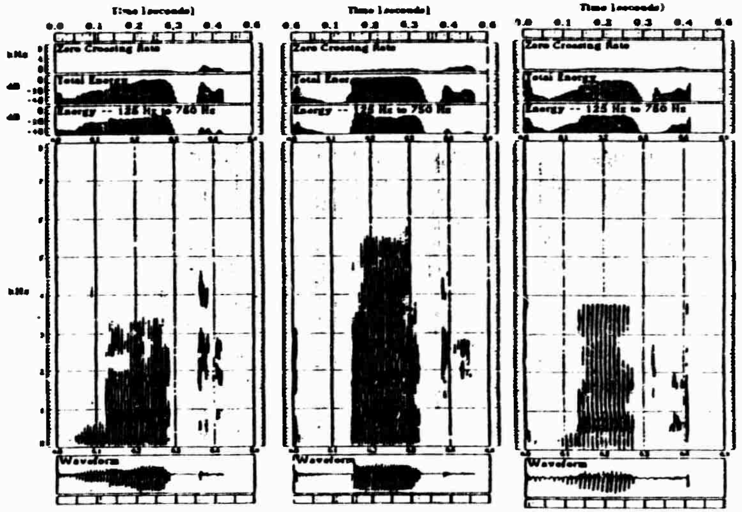


Figure 3.28: Spectrograms of the words *mack*, and *back*

Left: A spectrogram of the word *mack* spoken by a male speaker. Middle: A spectrogram of the word *back*, spoken by the same speaker. Right: A spectrogram of the word *mack*, spoken by a different male speaker

Another difficulty with a study of nasalized vowels is that different speakers nasalize to various degrees. Thus, one persons nasalized vowel could have the same characteristics as another's non-nasalized vowel. This phenomenon, also illustrated in figure 3.28, smears measurement distributions, and illustrates the difficulties associated with speaker-independent nasalized vowel identification.

For the acoustic analysis, there are a few procedures used to reduce the magnitude of this problem. First, the initial analysis is conducted on a speaker by speaker basis, to eliminate speaker-independent complications. In fact, in order to eliminate as much speaker and context variability as possible, the initial portion of the analysis is restricted to observing relative differences between minimal word pairs such as *skip* and *skimp*. This allows the observation of acoustic differences which are introduced by the presence of the nasal consonant.

A perceptual study, performed on a subset of the database, is also used to aid the analysis. Given a vowel token spliced out of the speech waveform, subjects must decide whether or not the vowel is next to a nasal consonant.[1] Once each vowel is given a nasality rating, acoustic characteristics may be correlated to establish a perceptual credibility. Of course, there are several issues which need to be considered in such a test, including whether or not untrained subjects know what a nasalized vowel is, or how natural the tokens are. However, the scores were found useful for guiding the initial study.

The inherent dynamic quality of nasalized vowel spectra also complicate the acoustic analysis. Unlike nasal consonants, nasalized vowels are not necessarily steady state sounds due to either the nature of nasalization (lowering or raising of the velum), or of the vowel itself (dipthongs). In either of these cases, the net effect is that the acoustic characteristics change with time. Averaging procedures, used throughout the analysis of the nasal consonants, are not adequate in this case. Figure 3.29 shows a spectrogram of the word *mode*, where the low frequency regions of the vowel are clearly changing with time.

[1] The perceptual study of the database is reported in detail in the next chapter.

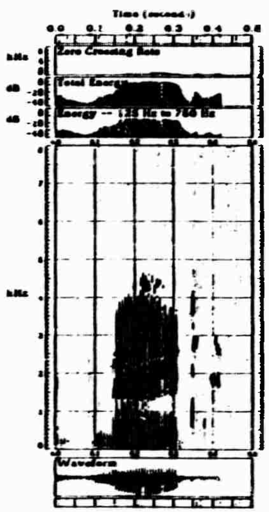


Figure 3.29: A Spectrogram of the word *made*

Of course, it is possible to try and track useful characteristics of the vowel (such as the resonance frequencies) for the duration of the vowel. This method was not used because such systems tend to be rather fragile, especially in nasalized vowels. Instead, the vowel was divided into *subsegments*, so that averaging procedures could be used in each subsegment to reduce measurement noise, yet changes between the different subsegments of the vowel, caused by increasing nasalization for example, would still be measureable. After some experimentation it was decided to use three subsegments in each vowel. Thus, whenever a measurement of some parameter was made on a vowel, there were three values returned. Each value represented an average of the parameter in one of the three, equally spaced, vowel subsegments.

The following sections report results of the study of the durational, and spectral characteristics of nasalized vowels.

3.2.1 A Study of Nasalized Vowel Duration

Sin.-e there are many contextual factors which can influence the duration of vowels, it would be unreasonable to expect to be able to distinguish nasalized vowels from non-nasalized vowels on the basis of duration alone. However, a minimal pair experiment was performed to establish if indeed there were any differences in duration. For this experiment, vowels in a nasal cunsonant context, such as *meat*, were paired with vowels in either a stop or fricative consonant cuntext, such as *beat*. The difference measure was calculated by subtracting the two vowel durations.

Figure 3.30 displays a histogram of the difference in duration for all of the vowel pairs. On average, vowels appear to be shortened by approximately 10 msec when they are put into a nasal consonant context. The spread of this distribution weakens the strength of this statement however. On closer inspection of the data, it appears that the greatest difference is between vowels in a fricative nasal cluster, such as *smock* versus *sock*, where the average difference is nearly 20 msec.

When the nasal consonant formed a post-vocalic cluster with a stop, or fricative consunant, the vowel duration was observed to vary with the voicing of the clustering consonant. Statistics for minimal pair duration differences between words such as *bend* and *bent*, or ones and once, may be found in figure 3.31. Vowels in a nasal stop consonant cluster, were observed tc be lengthened by 30 msec on average, when the stop cunsunant was voiced. Vowels in a nasal fricative consonant cluster, were observed to be lengthened by 10 msec on average, when the fricative cunsonant was voiced. Note that this durational change is much less significant than that observed in the nasal consonants in the same circumstances.

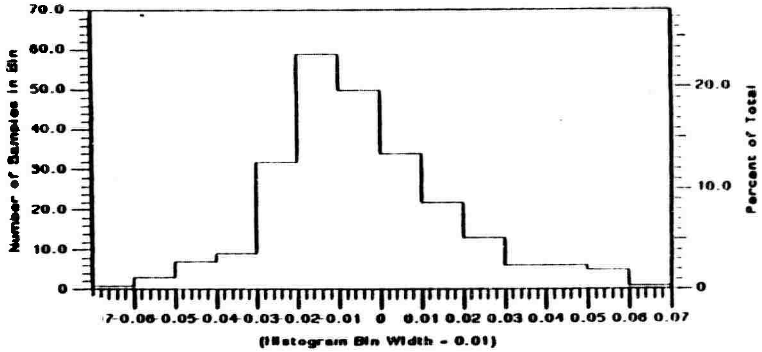


Figure 3.30: Vowel Duration Differences

The solid lines outline minimal pair duration differences between vowels in a nasal consonant context, such as the word *bent*, and those in a stop or fricative consonant context, such as the word *bet* (253 tokens).

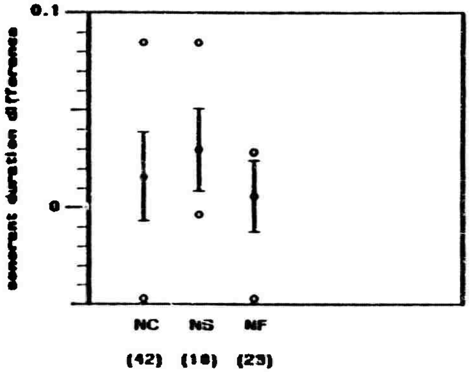


Figure 3.31: Vowel Duration Differences due to Voicing

This display summarizes the difference in vowel duration of minimal pairs in different voicing contexts. From left to right they are: all nasal consonant clusters (NC), nasal stop clusters (NS), nasal fricative clusters (NF). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

3.2.2 A Study of Nasalized Vowel Spectra

The spectral analysis of nasalized vowels was carried out in a manner similar to that of the nasal consonants. In the first stage of analysis, the goal was to establish differences between nasalized and non-nasalized vowels by comparing some form of average spectra. On the basis of these observations, general discriminating properties could be proposed and quantified using utterances of the database. The following sections describe the sequence of steps followed for the spectral analysis.

Spectral Averaging

Using the multiple spectra averaging technique described for the analysis of nasal consonants, statistics were collected for nasalized and non-nasalized vowels of each speaker. Initially, two average spectra were computed for each vowel (nasal context and non-nasal context). Figure 3.32 shows average spectra for an /æ/ for a male speaker.

Although there was a danger of smearing a significant amount of information by the averaging procedure, these plots were quite informative. The most noticeable difference between the nasalized and non-nasalized vowels was in the low frequency regions of the magnitude spectrum . On average, it was found that non-nasalized vowels had one resonance in the first formant region, while nasalized vowels had two. Of the two resonances found in the nasalized vowel, one could always be associated with a first formant. This resonance was labelled the "first resonance". The extra resonance, which could appear above, or below the first resonance, depending on the vowel height, was labelled the "nasal resonance", although it was clear that this resonance was not always a result of nasal coupling. In figure 3.33 for instance, which contains a nasalized, and non-nasalized /æ/ from the words *camp*, and *cap*, the first resonance is located at about 700 Hz, for both the nasalized and non-nasalized vowels. The nasal resonance is located near 250 Hz for both vowels as well. In figure 3.34 however, which contains a nasalized, and

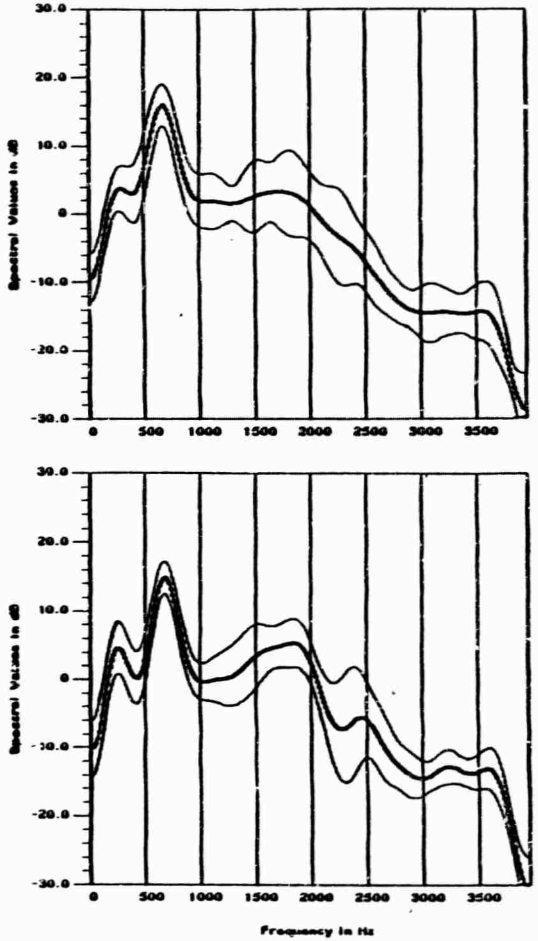


Figure 3.32: Average Spectral Shape of /æ/

The top display presents a statistical summary of the normalized, smoothed spectra of the non-nasalized /æ/ of a male speaker. The bottom display presents a summary of the nasalized /æ/ of the same speaker. The average spectral shape, shown by the dark line, is surrounded by lines which represent ooe standard deviation from the mean.

a non-nasalized /i/ from the words *technique*, and *beat*, the situation is different. Here the first resonance is located at about 350 Hz for both non-nasalized and nasalized vowels, and the extra resonance is located at 700 Hz.

In general it was observed that the nasal resonance would appear above the first resonance only for very high vowels, where the first resonance was centered below 400 Hz. Otherwise, the nasal resonance appeared below the first resonance.

Unfortunately, many non-nasalized vowels were observed to have a nasal resonance, as is clear from these figures.[2] This means that it is not always possible to distinguish nasalized from non-nasalized vowels by measuring the fraction of time that there is a nasal resonance in the vowel.

Fortunately, it was found that the nasal resonance was noticably more "distinct" in a nasalized vowel. There were two ways in which "distinctness" was manifested in the spectrum. First, the magnitude of the nasal resonance could iocrease relative to the first resonance. This could be caused by the first resonance decreasing in amplitude, or the nasal resonance increasing, or both. Second, the dip between the nasal resonance and the first resonance could deepen. Thus, if a non-nasalized and a nasalized vowel both happened to have an extra resonance, it is possible to discriminate between them by measuring the relative strength of the nasal resonance to the first formant. The previous two figures both provide good examples of how the nasal resonance is more distinct in the nasalized vowel.

Another observed characteristic of nasality was a smearing of the first resonance itself. In fact, when an extra resonance was not present, as was occasionaly observed in a nasalized vowel, a measure of the spread of energy about the first resonance was found to be the best indication of nasalization.

In summary then, by observation of spectra, a set of qualitative characteristics of vowel nasalization was proposed. Due to the variability of the environment, none

[2] The vowels produced by female speakers tended to have a low resonance in any context. This property was due to breathiness more than nasalization.

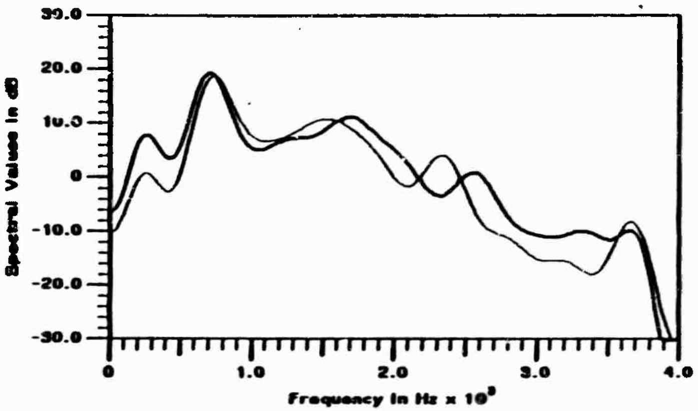


Figure 3.33: Overlay of Nasalized and Non-nasalized /æ/

This display contains spectra of the vowel /æ/ taken from the words *cap*, and *camp*. The light line is for the vowel in the non-nasalized context.

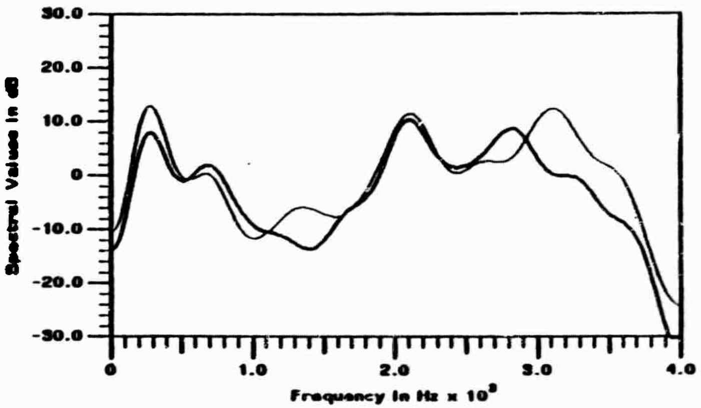


Figure 3.34: Overlay of Nasalized and Non-nasalized /i/

This display contains spectra of the vowel /i/ taken from the words *beat*, and *technique*. The light line is for the vowel in the non-nasalized context.

these characteristics was present in a nasalized vowel at all times. However, taken in combination, these properties were able to discriminate between nasalized and non-nasalized vowels. The next step was to quantify these observations. A set of algorithms was developed which were able to automatically extract these measures of nasalization. The details of the algorithms may be found in Appendix D. The following sections present the results of the quantitative analysis of nasalized vowels.

Minimal Pair Experiments

As a first step at quantifying the qualitative descriptions of nasalized vowels, the differences between the vowels of minimal word pairs such as *ben* and *bed*, *mack* and *back*, were observed. This procedure effectively eliminated speaker-dependent and vowel-dependent variability. The results of these minimal pair experiments have been summarized below.

The first parameter measured was a center of mass of the spectrum below 1000 Hz. A scatter plot of the average value of the center of mass in the middle portion of the vowel is shown in figure 3.35. The horizontal coordinate of a vowel is its value in a nasal environment, such as the /æ/ in *camp*. The vertical coordinate of the vowel is its pair value in a non-nasal environment. In this case the pair word is *cap*, spoken by the same speaker. Any vowel which has the exact same value of center of mass for both contexts will be located on the solid line. If a vowel has a lower center of mass value when it is nasalized, then it will lie above this line. If it has a higher center of mass when it is nasalized, the vowel appears below the line.

In general, it is clear that low vowels such as /æ/ tend to have a lower center of mass when they are nasalized, while high vowels such as /i/ tend to have very little change or a slight increase in center of mass. The change in values are a result in the increase in strength of the extra resonance produced through nasal coupling.

The next measure observed was the standard deviation of the local energy (within 500 Hz) around the center of mass. It was hypothesized that the low frequency energy of nasalized vowels is spread out, due to a weakening of the main formant and a strengthening of the extra resonance. Thus it would be expected that nasalized vowels would have a higher standard deviation than their non-nasalized counterparts. Figure 3.36 shows that in general this is true. The main exception to the rule is the vowel /æ/. This resulted from an artifact of the standard deviation computation which varied slightly with the center of mass value. The fact that /æ/'s have a much lower center of mass value when they are nasalized is enough to lower the deviation values slightly. Since the center of mass is relatively unchanged for other vowels, the standard deviation measure was not influenced to the same degree.

After observing general statistical properties of the low frequency spectra, measurements were made on the actual resonances. From observations of the average spectra, it was clear that nasalized vowels tended to have an extra resonance in the first formant region. Thus the first calculation measured the percentage of the time that there was an extra resonance in the vowel region. A similar calculation looked at the percentages in the three vowel subsegments. This measure was found to be more effective, since it allowed local areas of nasalization to stand out more than would be the case for an overall average. Figure 3.37 shows the value of the maximum percentage of the three subsegment (scaled by 100). In general, it may be seen that nasalized vowels nearly always have a greater maximum percent than their non-nasalized pairs. In fact, most of the nasalized vowels have a value greater than 0.8. Another distribution compares the values of the *minimum* percentage value of the three vowel subsegments, as shown in figure 3.38. Note that the only vowel which still has a high percentage is /æ/, indicating that this vowel nearly always has a low resonance.

Another observed quality of nasalization is the resonance dip, a measure of the drop in energy in between the two resonances, indicating the prominence of the weakest peak. Figure 3.34 plots the maximum value of this dip in the three subsegments of the vowel. Clearly nasalized vowels tend to have a larger dip than their counterparts. This observation strengthens the argument that the extra resonance becomes more distinct as nasalization increases.

The final me observed compared the relative difference in amplitude between the two resonances as shown in Figure 3.40, which plotted the minimum value of the difference for the three subsegments in the vowel. The resonance difference was calculated by subtracting the amplitude of the low resonance from the amplitude of the higher resonance. There are two points to note here. In low vowels, the extra resonance appears below the first formant. Thus the difference value will tend to be positive. As the vowel becomes more nasalized the extra resonance becomes stronger so the difference becomes smaller. In some cases, the extra resonance becomes so large that the difference becomes negative. The exact opposite is true of high vowels when the extra resonance appears above the first formant. In this case the difference starts oft negative and, as the extra resonance grows in magnitude, becomes more positive. In the extreme case (never observed), this resonance would be larger than the first resonance, making the difference positive. Thus, the effect of nasalization on the resonance difference depends on the vowel height.

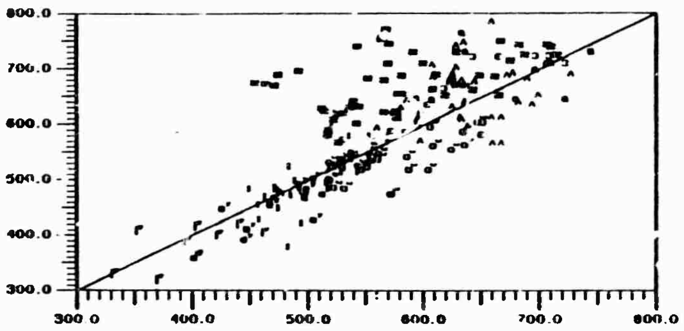


Figure 3.35: Scatter Plot of Center of Mass

This display indicates relative differences in center of mass between nasalized vowels and their non-nasalized counterparts. The horizontal coordinate of a vowel is its value in a nasal context (such as the /ɛ/ in *bent*). The vertical coordinate of the vowel is its value in a similar, but non-nasal, context (such as the /ɛ/ in *bet*).

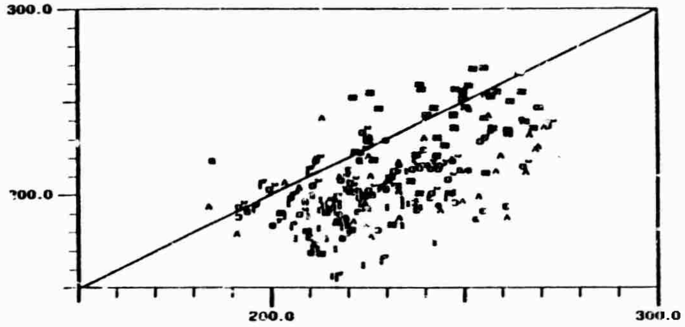


Figure 3.36: Scatter Plot of Standard Deviation

This display indicates relative differences in standard deviation between nasalized vowels and their non-nasalized counterparts. The horizontal coordinate of a vowel is its value in a nasal context. The vertical coordinate of the vowel is its value in a similar, but non-nasal, context.

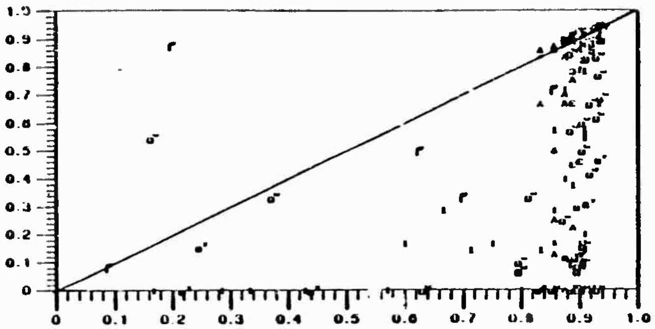


Figure 3.37: Scatter Plot of Maximum Percent

This display indicates relative differences in maximum percentage between nasalized vowels and their non-nasalized counterparts. The horizontal coordinate of a vowel is its value in a nasal context. The vertical coordinate of the vowel is its value in a similar, but nasal, context.

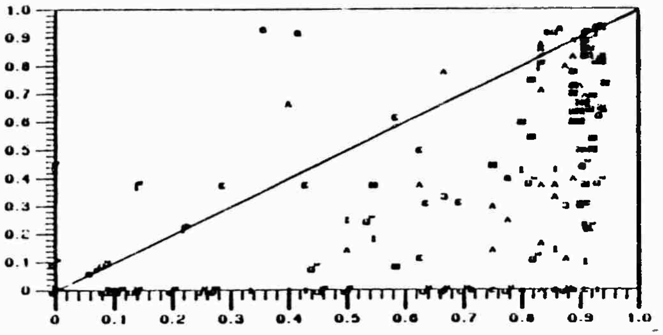


Figure 3.38: Scatter Plot of Minimum Percent

This display indicates relative differences in minimum percentage between nasalized vowels and their non-nasalized counterparts. The horizontal coordinate of a vowel is its value in a nasal context. The vertical coordinate of the vowel is its value in a similar, but non-nasal, context.

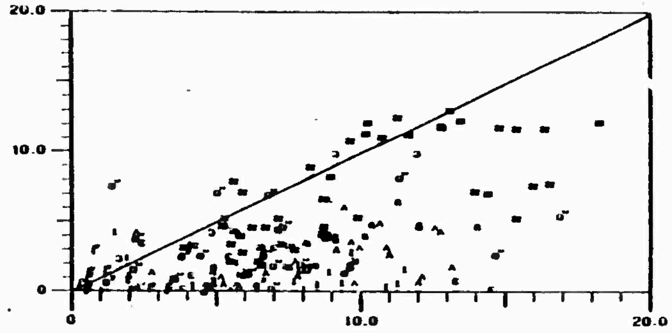


Figure 3.39: Scatter Plot of Maximum Resonance Dip

This display indicates relative differences in resonance dip between nasalized vowels and their non-nasalized counterparts. The horizontal coordinate of a vowel is its value in a nasal context. The vertical coordinate of the vowel is its value in a similar, but non-nasal, context.

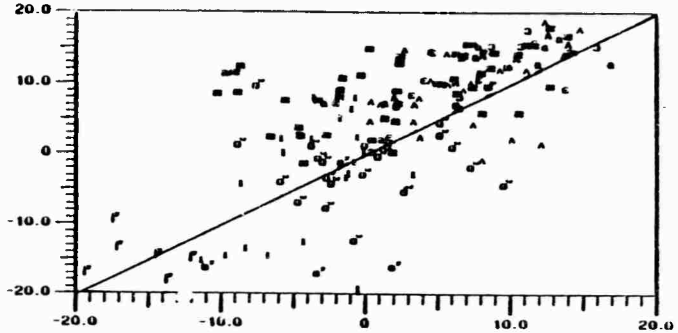


Figure 3.40: Scatter Plot of Minimum Resonance Difference

This display indicates relative differences in resonance difference between nasalized vowels and their non-nasalized counterparts. The horizontal coordinate of a vowel is its value in a nasal context. The vertical coordinate of the vowel is its value in a similar, but non-nasal, context.

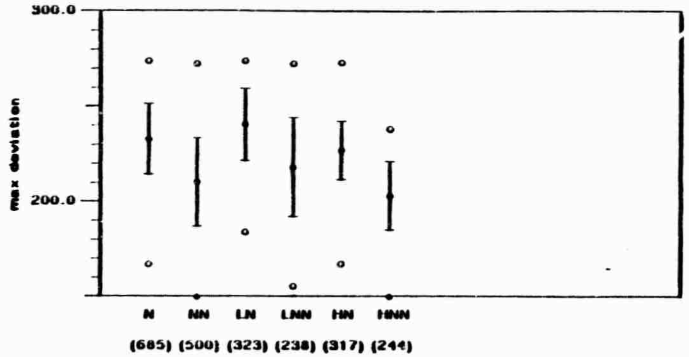


Figure 3.42: Statistics of Maximum Standard Deviation

This display summarizes the standard deviation of nasalized and non-nasalized vowels in different contexts. From left to right they are: all nasalized vowels (N), all non-nasalized vowels (NN), nasalized low vowels (LN), non-nasalized low vowels (LNN), nasalized high vowels (HN), and non-nasalized high vowels (HNN). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

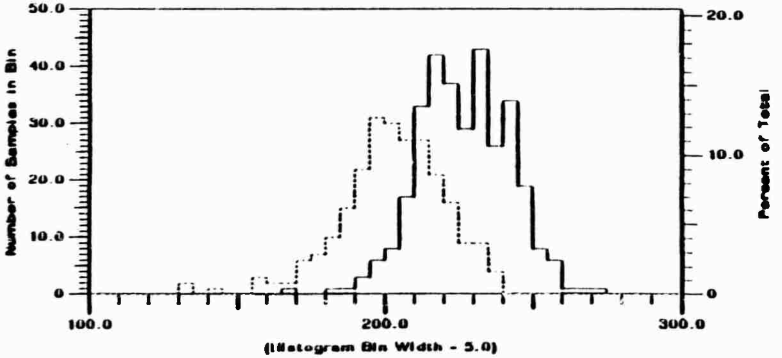


Figure 3.43: Histogram of Standard Deviation of High Vowels

This display contains a histogram of the standard deviation of all high vowels. The dark lines are the distributions of nasalized vowels (317 samples). The dashed lines are the distributions of non-nasalized vowels (244 samples). Values are in Hz.

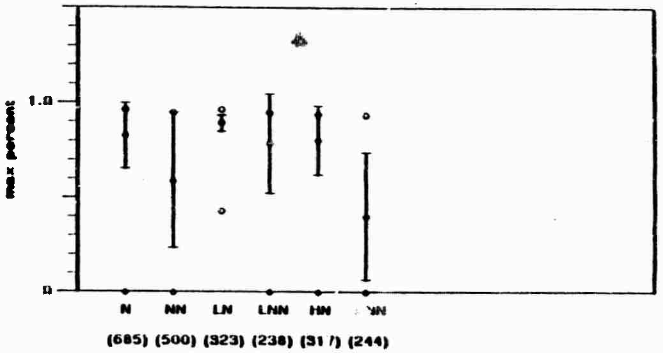


Figure 3.44: Statistics of Maximum Percent

This display summarizes the maximum percentage of nasalized and non-nasalized vowels in different contexts. From left to right they are: all nasalized vowels (N), all non-nasalized vowels (NN), uasalized low vowels (LN), non-nasalized low vowels (LNN), nasalized high vowels (HN), and non-nasalized high vowels (HNN). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The number of samples in each context are indicated below the display.

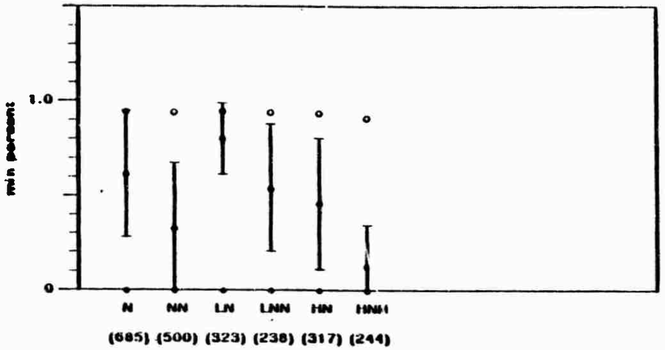


Figure 3.45: Statistics of Minimum Percent

This display summarizes the minimum percentage of nasalized and non-nasalized vowels in different contexts, which are the same as those in figure 3.44. The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The numer of samples in each context are indicated below the display.

General Results

Once minimal pair experiments had established some relative results, distributions were made for all of the vowels. It was found useful to retain a high-low distinction in the distributions however, since low vowels tended to have a more distiuct nasal resonance than did high vowels. The resulis of these experiments have heen summarized in the following paragraphs.

With the exception of the vowel /æ/, center of mass was not at all effective in discriminating nasalized from non-nasalized vowels.[3] Figure 3.41 shows that center of mass is effective in separating high vowels from low vowels, so this parameter could he of use if vowel height was unknown.

As may he seen in figure 3.42, standard deviation was quite effective in distinguishing nasalized from non-nasalized vowels. In general, nasalized vowels have a higher standard deviation value than non-nasalized vowels. Among each vowel type (high or low), standard deviation does quite well at separating the two groups. Figure 3.43 illustrates the distrihutions for high vowels.

Figures 3.44 and 3.45 show that the extra resonance percentage measure is also effective in separating nasalized and non-nasalized vowels. From the statistics of the maximum percent region, it is clear that nasalized vowels will always have a high percent value (especially low vowels), while many non-nasalized vowels will not. The minimum percent region shows that low vowels have a resonance throughout the vowel. This is not the case for high vowels, which have a smaller minimum percent. However, since non-nasalized high vowels have even smaller values, this calculation is a good discrimination measure.

Figure 3.46 displays the statistics of the resonance dip measure. Although this calculation is clearly useful, it points out the necessity of being able to differentiate between high vowels and low vowels, since non-nasalized low vowels have a very similar distribution to nasalized high vowels.

The statistical distributions of the measure of difference, shown in figure 3.47, are perhaps the most difficult to interpret since they appear to overlap. The idea behind this measure was that as the extra resonance hecame stronger, the difference between it and the first resonance would get smaller. Thus we would expect that as a vowel hecomes more nasalized the resonance differeuce will go to zero. This was certainly true for the low vowels. Unfortunately, there was a problem with computing this for high vowels, since for high back vowels such as /u/, or /o/, the second formant could get confused with a possible low re ce. This resulted in the distribution being rather spread out for non-nasalized high vowels, since there were some very negative values and other very positive values.

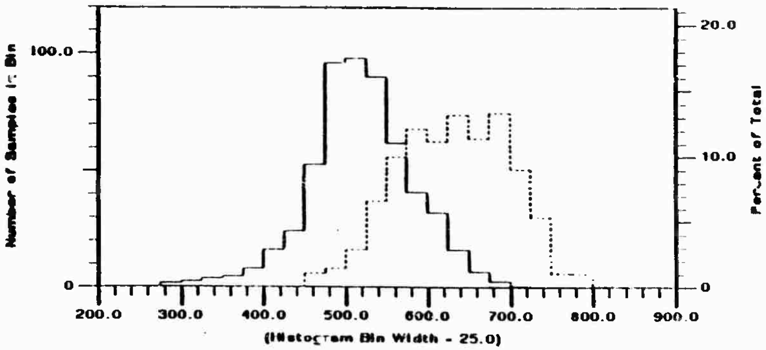


Figure 3.41: Histogram of Center of Mass

This display contains a histogram of the center of mass of all vowels. The dark lines are the distributions of the high vowels (56t tokens). The dashed lines are the distributions of low vowels (56t tokens). Values are in Hz.

[3] In fact, the change in height of a nasalized /æ/ may be influenced more by phonological rules of American English than by acoustic changes due to nasal coupling [32]

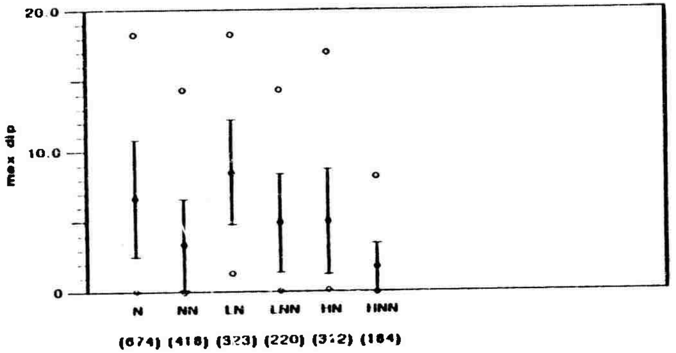


Figure 3.46: Statistics of Maximum Resonance Dip

This display summarizes the maximum resonance dip of nasalized and non-nasalized vowels in different contexts. From left to right they are: all nasalized vowels (N), all non-nasalized vowels (NN), nasalized low vowels (LN), non-nasalized low vowels (LNN), nasalized high vowels (HN), and non-nasalized high vowels (HNN). The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The numer of samples in each context are indicated below the display

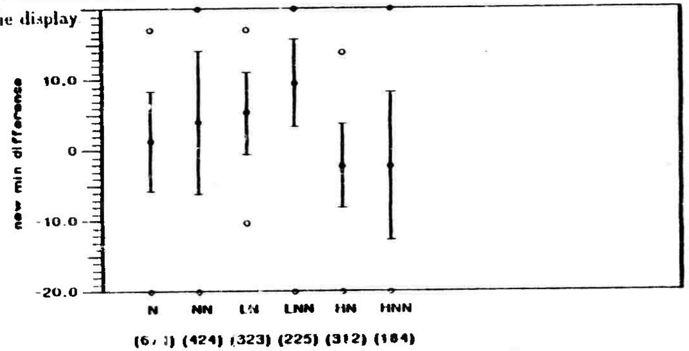


Figure 3.47: Statistics of Minimum Resonance Difference

This display summarizes the minimum resonance difference of nasalized and non-nasalized vowels in different contexts, which are the same as those in figure 3.46. The average value is indicated by a filled circle. The vertical lines indicate one standard deviation, and the open circles display the maximum and minimum values. The numer of samples in each context are indicated below the display

Discussion

The variability of vowel spectral shapes hindered the study of nasalization. Since the main area of interest was in the first resonance region, the difficulties lay with making sure that the second formant never influenced the computations. This was naturally difficult if analysis ranges were to be kept high enough to include all of the first formants of low vowels, and kept low enough to exclude all second formants of back vowels. Since this boundary is not a fixed threshold, there are bound to be some cases where the measurement algorithms do not work correctly, as has been pointed out.

In spite of these tokens, whose main contribution was to add noise to the distributions, the acoustic study established several useful measures of nasality. The most robust measnre of nasalization is the addition of an extra resonance in the low frequency region. As a result, energy in the first resonance region is more spread out, as indicated by the measure of standard deviation.

Also apparent from the acoustic study is that it is possible to discern relative degrees of nasalization by measuring the strength of the extra resonance frequency relative to the first resonance, and by measuring the amount of time that it is present in the vowel.

These observations are consistent with those made by other researchers in the past [20], [21], who have noted the presence of an extra resonance in nasalized vowels. The low resonance has been typically measured around 250 Hz. For high vowels such as /i/, the extra resonance has been observed to be located above the first formant, as was the case in this analysis. Hawkins and Stevens have suggested that nasal coupling introduces a pole zero pair into the first resonance region, and suggest that it is the zero which is the main cause of amplitude reduction of the first formant. The presence of a zero in between the low resonance and the first formant was also noted by Hattori et al., and might explain the effectiveness of the spectral dip measure of this analysis.

There are other, secondary characteristics of nasality which have been reported previously and which have not been quantified in this analysis. Most of these properties, such as observations of higher formant motion, or additional nasal resonances, were either not observed in this data, or where too difficult to attempt to extract automatically. For instance, quite often, high front vowels will exhibit another extra resonance in between the first two formants. Extracting this property however, would probably require some form of formant tracking, a difficult task in itself. Thus, this characteristic was not quantified.

Finally, it is worth examining the fact that female vowels exhibit a low resonance irrespective of context. Since these vowels are not all nasalized, there must be some other explanation for their presence. A previous study of vowels has shown that the first harmonic is enhanced when vowels have a breathy quality [2]. Since the pitch of female speakers is quite often found in the 200 to 300 Hz range, the low resonance could easily be a measure of breathiness in female speech. Although the presence of this resonance reduces the usefulness of the percentage measure for female speakers, identification of nasalized vowels is still possible, since the low resonance is strengthened when a vowel is nasalized.

3.3 Chapter Summary

The following points were established in this chapter:

1. The most robust acoustic property of a nasal consonant is a steady, low frequency resonance, which dominates the spectrum. The resonance is characterized by temporal and spectral stabilty, and by its local relative strength, properties which were quantified by the measure of low resonance percentage, and low resonance height, respectively.

2. The most robust acoustic property of a nasalized vowel is the presence of an extra resonance in the first formant region. Depending on the height of the vowel, the extra resonance may appear above, or below the first formant. Even if the extra resonance may not be resolved from the first formant, the first resonance region is more spread out when a vowel is nasalized, a property which was quantified by the measure of standard deviation.

3. It is possible to discern relative degrees of nasalization by measuring the strength of the extra resonance relative to the first formant, and by measuring the amount of time that it is present in the vowel.

Chapter 4

Recognition Experiments

After observing the acoustic characteristics of nasal consonants and nasalized vowels, preliminary investigations were initiated to evaluate the potential use of these properties in speech recognition. A detailed description of the experiments that were conducted on nasal consonant and nasalized vowel detection are presented in this chapter. These experiments cannot realistically simulate a true test environment since the evaluations were made on the same database as the acoustic study. However, they do provide an indication of their potential for use in speaker-independent, speech recognition systems.

4.1 The Task

There are many different ways of restricting the problem of speaker-independent, continuous-speech, automatic nasal consonant recognition. Since the acoustic measures developed in the acoustic study were designed for the discrimination, the recognition task was structured as an identification problem. In a typical scenario, the nasal consonant detection system is given a test token and training data. The system must then classify the token as either a nasal consonant or an impostor sound. The nasalized vowel detection system must classify a test token as either next to a nasal consonant (nasalized), or not next to a nasal consonant

(non-nasalized). Note that the evaluation procedure of the nasalized vowel detection system is not a true judge of nasalization, since some vowels in a non-nasal context will be nasalized, while some vowels in a nasal context will hardly be nasalized at all. A better evaluation measure would be to compare system decisions with those of human listeners.

Structuring the task in this format simplifies the problem, since it eliminates the need to detect the boundaries of the nasal consonant or vowel. These systems might be considered as specialized modules of a recognition system which are called upon only in situations which require their expertise.

4.2 The Strategy

The acoustic study quantified several parameters which characterize nasal consonants and nasalized vowels, and may discriminate them from similar sounds. Thus, it is reasonable to incorporate these measurements into detection systems for the task in hand. A given test token is then associated with a set of n values, corresponding to a set of acoustic measurements made on the test token. If we consider the set of values as a vector in an n-dimensional space, we are faced with a multidimensional decision making problem.

Multiva ate decision making becomes a straightforward process when each of the parameters involved may be assumed to have jointly Gaussian distributions. In these cases, decision making is reduced to finding the distance from the test token point, to each of the normalized distributions of the possible candidates

$$D_i = (\vec{X} - \vec{m}_i)^T C_i^{-1} (\vec{X} - \vec{m}_i) \qquad (4.1)$$

where D_i is the distance from the test token to candidate i; $\vec{m}_i$ is the mean vector of the parameters in the ith distribution; C_i is the covariance matrix of the parameters in the ith distribution; and $\vec{X}$ is a vector of the parameter values of the test token. For nasal consonant detection, the two candidate distributions are nasal consonants or impostors sounds, and so two distances are computed. Candidate membership is dictated by the minimum distance value.

Gaussian discrimination techniques are popular since they are quite simple, and the distance metrics correspond to maximum likelihood decisions [48]. Further, many parameter distributions may be reasonably approximated by a Gaussian of some form. When this is not the case, it is quite often possible to transform the distribution (by taking logs for example) so that a Gaussian approximation becomes reasonable. The question of a joint distribution is more difficult to account for, unless the parameters can be shown to be statistically independent.

When the joint Gaussian distribution assumption is not valid, some other procedure might be superior. Another approach which is often used, is a binary tree classifier, where, at each node in the tree, a split is made according to some criterion in one of the parameters [5]. In this fashion, a tree may be constructed which separates the data into categories without making assumptions about the underlying distributions of the parameters. Decisions are made at the bottom level of the tree based on a majority rule of the training data. Thus, if a test token happens to end up in a slot where 20 tokens of type A and 5 tokens of type B were observed during training, the test token would be classified as type A with a score of 0.8.

Although the tree classifier is attractive in the sense that it makes no assumptions of underlying distributions, it suffers from the fact that decision thresholds and actual tree structures can vary substantially, depending on the training data used.

An alternative way to combine the data would be to evaluate each parameter individually, and combine the scores at the last stage to establish some overall decision. For a binary decision (nasal or impostor), each parameter need only return a single value such as the log ratio of the likelihood that the token is nasal to the likelihood that the token is an impostor. This technique eliminates any potential multivariate information, and unless the parameters were statistically independent, might be expected to perform poorly. However, when the parameter distributions are inappropriate for standard Gaussian techniques, the likelihood approach was found to be more effective.

All of these approaches require some form of *a priori* knowledge of the distributions of the parameters. These are established through the use of training data provided to the systems. In the Gaussian approach, these values are used to compute means and covariances. In the binary classifier approach, they are used to establish node thresholds. In the the likelihood procedure, distributions are created so that an incoming test sample may be accorded a likelihood value. The actual distributions used were simple normalized histograms of the measurements. Bin widths of the histograms were set manually to ensure that the distributions would be reasonably shaped.

4.3 The Experiment

Systems were evaluated using the utterances of the database. For the nasal consonant detection task, data was divided into two groups, those in the nasal consonant class, and those which might be confused with nasal consonants (called the impostor class). These sounds included any phoneme which had acoustic characteristics similar to nasal consonants such as voice bars, liquids and glides, or weak voiced fricatives. For evaluation, there were 520 nasal consonant tokens, and 695 impostor tokens which included 357 semivowels, and 338 voice bars.

For the nasalized vowel task, the data was divided into similar groups, with the exception being that the test token was the vowel adjacent to either a nasal consonant or an impostor sound. For evaluation, there were 685 "nasalized" vowels and 500 "non-nasalized" vewels.

Ideally, the choice of the impostor sounds should be governed by a knowledge of perceptual errors. However, studies which have examined perceptual confusions between individual phonemes are scarce. Miller and Nicely have examined perceptual confusions in noisy,and band-limited signals [49]. While the study is of interest, since it indicates that nasal consonants are indeed confused with liquids, glides, and voiced stops, it is not possible to make a strong case for using their results, since the test environments are quite different. Therefore, the criterion for choosing impostors was governed more by acoustic similarity and recognition difficulty (from reports of other recognition systems).

Systems were evaluated using a rotational procedure. In each step, systems were allowed to train on the data from five of the six speakers in the database, and were tested on the data from the sixth speaker. This approach is the best approximation to a speaker-independent task, given the limited amount of data available. The following sections report the results of nasal consonant and nasalized vowel detection.

4.3.1 Detection of Nasal Consonants

There were five measures from the acoustic study which were incorporated into the nasal consonant detection system. These included:

1. *Total Energy*. The average amount of energy in the token.
2. *Energy Stability*. The average amount of change in energy in the middle of the token.
3. *Low Resonance Percentage*. The percentage of the time that there was a low frequency resonance below 350 Hz in the token.
4. *Low Resonance Amplitude*. The average amount of energy in the low frequency regions relative to total energy in the token.
5. *Low Resonance Height*. The average energy drop from the low frequency resonance to the regions immediately above.

Since it was unclear as to which decision strategy would yield the best results, an initial analysis was conducted to determine which of the three methods discussed previously performed the best. For simplicity, the systems were allowed to train on all of the tokens, since the goal was to measure the relative performance of the different strategies.

For the first examination of the data, a standard Gaussian technique was employed. Evaluating the data on the nasal consonants and impostors yielded a correct identification rate of 79%. The fact that this simple approach did so well was actually surprising, since many impostor distributions were non-Gaussian. Observation of these distributions indicated that many of the bi-modal distributions were effectively a sum of two rather standard distributions, one consisting mainly of voice bars, and the other of semivowels. This observation was also made by Mermelstein [48].

In an attempt to remedy this situation, the procedure was modified by separating the voice bars from the semivowels so that there were actually two impostor groups. An incoming test token would compute three distances, instead of two. If the minimum distance was to either of the impostor distributions, the token was labeled an impostor. Otherwise the test token was called a nasal. The average correct detection rate for this modified approach improved to 85%. It is of interest to note that glides and voice bars were rarely confused with each other, and that most of the errors were caused by labeling the nasal consonants impostors.

A binary tree classifier was evaluated next. Testing the binary tree classifier on the same data used for training is unfair, since it is possible to grow the tree during training, until there is but one element in each branch. Testing on the same data will naturally result in 100% accuracy However, by restricting the tree to depths of around four nodes, detection rates of 91% were obtained. In order to test the sensitivity of the node thresholds, the tree was allowed to train on half of the data, and was tested on the other half (speakers were still mixed). In this case, the performance declined to 87%, indicating that thresholds were slightly sensitive to the data.

The final evaluation procedure summed the set of individual log likelihoods to come up with an overall nasal likelihood score. An average score of 89% was obtained. Confusions for all three approaches are summarized in table 4.1.

Table 4.1: Nasal Consonant Detection Confusions

	Gaussian		Tree		Likelihood	
	Nasal	Impostor	Nasal	Impostor	Nasal	Impostor
Nasal	70	30	86	14	94	6
Impostor	7	93	12	88	16	84

Since the log likelihood strategy performed slightly better than any other, and appeared to be a quite robust, it was evaluated again with the circular evaluation procedure described previously. For this case, the detection rate dropped to 88% (half a percentage point). The lack of significant decrease in the detection rate is encouraging, since it indicates that the acoustic parameters being extracted are reasonably speaker independent.

Discussion

Comparisons to other nasal consonant recognition systems are not valid at this stage of analysis, since the evaluation took place on the same database as the acoustic study. Once these parameters are tested on completely different database, the results will provide a better estimate of the speaker-independent capabilities of the system. Apart from this large qualification, it should be noted that there have not been many speaker-independent evaluations reported in the literature. Mermelstein probably had one of the more successful recognition systems although he trained and tested on only two male speakers [48].

From a recognition standpoint, it would be useful to establish the contribution

made by each parameter to the overall decision. Indications from the binary tree classifier were, that the percentage measure was the most valuable, followed by the measure of the low resonance height. This implies that the main property of nasal consonant which distinguishes them from other sounds, is a continuous, low frequency resonance which dominates the spectra below 1000 Hz.

From a perceptual perspective, it would be interesting to know if the decisions made by this system are related at all to what humans would do given the same task. This topic is pursued further later on in this chapter.

4.3.2 Detection of Nasalized Vowels

There were six measures from the acoustic study which were incorporated into the nasalized vowel detection system. These included:

1. *Center of Mass*. The average value of the center of mass in the middle of the token.
2. *Standard Deviation*. The maximum value of the average standard deviation in the three vowel subregions.
3. *Maximum Resonance Percentage*. The maximum percentage of the time there is an extra resonance in the three vowel subregions.
4. *Minimum Resonance Percentage*. The minimum percentage of the time there is an extra resonance in the three vowel subregions.
5. *Maximum Resonance Dip*. The maximum value of the average dip between the first resonance and the extra resonance in the three vowel subregions.
6. *Minimum Resonance Difference*. The minimum value of the average difference between the first resonance and the extra resonance in the three vowel subregions.

As was the case for the nasal consonants, an initial analysis was performed to determine which of the three methods performed the best. Once again for simplicity, the systems were allowed to train on all of the tokens.

Using the standard Gaussian technique, a correct detection rate of 71% was obtained. No further progress was made with this technique, since there were no obvious ways to divide the data, as was the case for the nasal consonants.

When the binary tree classifier was trained and tested on the same data set, a nasalized vowel detection rate of 84% was achieved. However, when the tree was trained on half of the data, and tested on the other half, the correct detection rate fell to 79%, indicating that the node thresholds were quite sensitive to the data.

Using the log likelihood procedure, an average score of 78% was obtained. Confusions for all three approaches are summarized in table 4.2.

Table 4.2: Nasalized Vowel Detection Confusions

	Gaussian		Tree		Likelihood	
	Nasal	Non-nasal	Nasal	Non-nasal	Nasal	Non-nasal
Nasal	59	41	90	10	84	16
Non-nasal	9	91	36	64	30	70

Although the log likelihood procedure did not perform quite as well as the binary tree classifier in the initial analysis, it was used for the speaker independent test because it was found to be more stable than the tree classifier. Using the circular evaluation procedure, an average detection rate of 74% was obtained.

Unlike the nasal consonant detection systems, which performed uniformly across different speakers, and phonetic contexts, the performance of the nasalized vowel detection system varied substantially with the environment. In order to measure the difficulty of different contexts, a series of smaller evaluations, using the log likelihood procedure, were made on subsets of the vowels. The results of these experiments are summarized in table 4.3.

Table 4.3: Nasalized Vowel Detection

Evaluation	Detection Rate		
	Nasalized	Non Nasalized	Average
All	81	67	74
Male	83	78	81
Female	66	60	63
High	82	75	79
Low	75	63	69
Male High	82	75	79
Male Low	88	83	85
Female High	74	71	73
Female Low	56	67	61

From this data, it is clear that discrimination between nasalized and non-nasalized low vowels, spoken by female speakers, is quite difficult. It is also evident, that it is more difficult to detect nasality in the vowels of female speakers than in those of male speakers. Note that care must be taken to interpret the last four entries of the table since there were only two speakers in the training distributions.

Discussion

While 74% correct is better than chance, it leaves a large number of vowels for which no confident statement may be made about the presence of an adjacent nasal consonant. The main reason for this is that speakers nasalize to different degrees. Thus, in attempting to operate in a speaker independent environment, the individual distributions are being smeared.[1]

The deterioration of the detection scores for female speech is understandable, since there is often an extra low resonance in the sonorant regions, as illustrated for the vowel /æ/, in figure 4.1. Since, an extra resonance is a major acoustic

[1] Some earlier speaker dependent studies obtained detection rates over 10% better than those reported here.

difference between uasalized and non-nasalized vowels, it is natural to expect system performances to deteriorate when the low resonance is present in the speech signal irrespective of nasality. It is interesting to note that for female speakers, the system was able to identify nasalization in high vowels better than for low vowels. Since the low resonance of female speakers is always below the first formant, high vowels which have a nasal resonance *above* the first formant are uniquely nasal, and so, may be identified correctly.

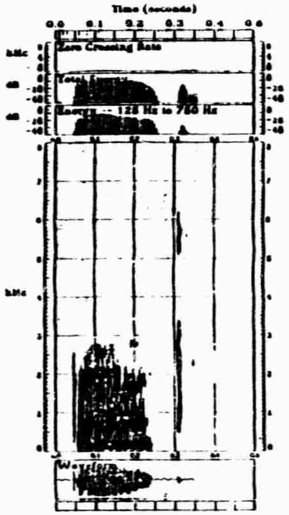


Figure 4.1: A Spectrogram of the word *back*

The performance of male speakers, is more intuitively acceptable, since the nasal resonance tends to be more "distinct" in low vowels, than in high vowels. Thus, one would expect to be able to detect nasalization more successfully in low vowels, which was confirmed in this experiment.

Of course, it is not clear how well the system is actually measuring nasalization in vowels. One way to get a better idea of this would be to perform a perceptual experiment on listeners given the same task. If the detection systems were extracting some perceptually relevant property of nasalization, then the there should some correlation between the two measures. The next section investigates this concept in more detail.

4.4 A Perceptual Evaluation

In order to provide a perceptual evaluation of the automatic detection systems , a listening experiment was performed which tried to measure people's ability to perceive nasality when all context had been stripped away. The experiment consisted of tests in which part of the speech waveform was extracted from continuous speech. For the three tests, the speech segment corresponded to:

- a murmur such as a nasal consonant, glide, or voice bar,
- a vowel adjacent to a murmur and,
- both the murmur and the adjacent vowel.

Each test consisted of forty tokens (twenty nasal and twenty non-nasal) spliced from utterances in the data base. Each token was smoothed at the ends to eliminate artifacts due to the splicing procedure, and played three times in succession. Subjects were asked to decide whether they thought the token contained a nasal (or for the second test, if the vowel was adjacent to a nasal) or a different speech sound.

The results from a panel of 20 listeners indicate that nasal consonants can be identified correctly about 65% of the time. There is some dependence on the duration of the segment but not a significant amount. Listeners were able to tell nearly 65% of the time whether a vowel was adjacent to a nasal consonant or not. Low vowels tended to be called nasal irrespectively of the presence of a nasal consonant. Listeners performed the hest when they were given both the murmur and the adjacent vowel to listen to, scoring over 85%.

Comparison to Detection Systems

When the tokens of the first listening test were run on the nasal consonant detection system, 64% of the tokens were identified correctly. This was effectively the same as the listeners. The nasality scores produced by listeners and the detection system seem to he rather correlated, as shown in figure 4.2.

When the tokens of the second test were run on the nasalized vowel detection system, 74% of the tokens were correctly identified. This result is notably hetter than that obtained hy human listeners. For this test as well, there were indications that the scores were somewhat correlated, as shown in figure 4.3.

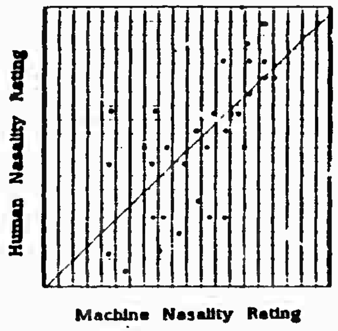


Figure 4.2: Nasality Rating for Murmur Tokens: Human versus Machine

This figure plots two measures of nasality for murmur tokens in the perceptual study. On the vertical axis are listeners nasality rating of the token. Likelihood scores given by machine are plotted on the horizontal axis.

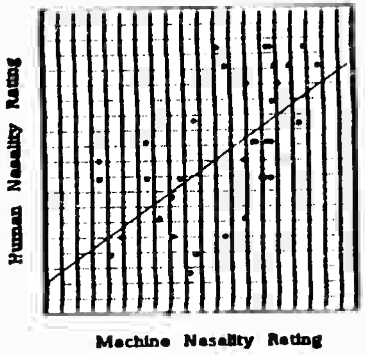


Figure 4.3: Nasality Rating for Vowel Tokens: Human versus Machine

This figure plots two measures of nasality for vowel tokens in the perceptual study. On the vertical axis are listeners nasality rating of the token. Likelihood scores given by machine are plotted on the horizontal axis.

Discussion

Since the nasal consonant detection system scored well below average on the murmur tokens used in the perceptual experiment, it is clear that these sounds were a difficult subset of the database. Thus, in general, one could expect listeners to be more successful at the nasal consonant detection task than was observed here. The fact that the results were somewhat correlated provides an indication that the nasal consonant detector is extracting relevant parameters from the speech signal.

Perhaps one of the more informative results of the vowel study was that it indicated that listeners are indeed able to use information in the vowel to detect nasal consonants. These results support the hypothesis of Ali et al that listeners use nasalization to lighten the phoneme processing load [1]. The fact that the vowel detection system performed better than listeners at this task is probably due in part to the fact that nasalized vowels have no phonetic distinction in American

English. Thus, untrained listeners were not aware of the concept of nasalization, and had a harder time detecting this property. Another reason for this difference in performance could have been that the detection system was allowed to train on utterances spoken by the same speakers, while human listeners were not.

4.5 Chapter Summary

The main points of this chapter are:

1. Using a log likelihood decision strategy employing robust measures established in the acoustic analysis, nasal consonant detection rates of 88% were obtained.
2. Using a similar decision strategy, a nasalized vowel detection rate of 74% was obtained. Detection rates varied substantially with speaker sex, and vowel height. The best decision rate of 85%, was obtained for low vowels spoken by male speakers. The worst decision rate of 61%, was obtained for low vowels spoken by female speakers.
3. A perceptual evaluation of a subset of the database indicates that system decisions tend to be correlated with decisions made by human listeners performing a similar task.

Chapter 5

Summary and Future Work

5.1 Summary

There are several conclusions which can be made from this research. First, the acoustic analysis established that nasal consonants are characterized by a low resonance, typically centered between 200 and 350 Hz, which dominates the overall spectrum. Another property of the low resonance, which was found to be typical of nasal consonants, was a sudden drop in energy at slightly higher frequencies in the first resonance region. This measure was found to be most effective in discriminating nasal consonants from semivowels. This parameter should also rule out most vowels, with the exception of some high front vowels such as /i/, or a raised /u/.

The acoustic analysis also found that the most robust measure of nasalization is the presence of an extra resonance in the low frequency region, resulting in a first resonance region where the energy is more spread out, as indicated by the measure of standard deviation. The acoustic study also established that it is possible to discern relative degrees of nasalization by measuring the relative strength of the extra resonance to the first resonance, and by measuring the amount of time that it is present in the vowel.

Finally, the preliminary investigations of nasal consonant and nasalized vowel detection provide indications that these acoustic properties are useful for applications in speaker-independent speech recognition systems.

5.2 Future Work

Although this research observed many characteristics of nasality, there are still many areas which require further investigation. One area which was only briefly examined, was the transition region between the nasal consonant and the adjacent vowel. This time interval is worthy of serious study, since it contains the most information about the place of articulation of the nasal consonants. In addition, it contains pertinent information for discriminating semivowels from consonants.

As illustrated in figure 5.1 for the word *need*, the transition region of prevocalic nasal coosonants is denoted by a sudden spectral change at high frequencies, with limited formant transitions in the vowel. This information can be used to discriminate nasal consonants from semivowels. Figure 5.1 also shows an /l/ from the word *lead*, which, although having similar acoustic characteristics to a nasal consonant, may be eliminated as a potential nasal due to a lengthy second formant transition.

Another characteristic worth quantifying, is the extension of the low frequency resonance into an adjacent vowel, as has been illustrated before in figures 2.6, 3.28, and 3.29. Although this property is not apparent for all vowels, it is a very powerful indication of the presence of a nasal consonant when it exits.

From a speech recognition standpoint, there are several ways in which this work could be extended. First, it is clear that the evaluation performed in this work is inadequate since it it was based on the same database as the acoustic analysis. For a true evaluation of the parameters developed in this work, a totally different database should be used. For speaker-independence, the database should have a large number of speakers. Since the acoustic measurements developed in the acoustic study are proposed for continuous speech, it also would be appropriate to collect a sentence database.

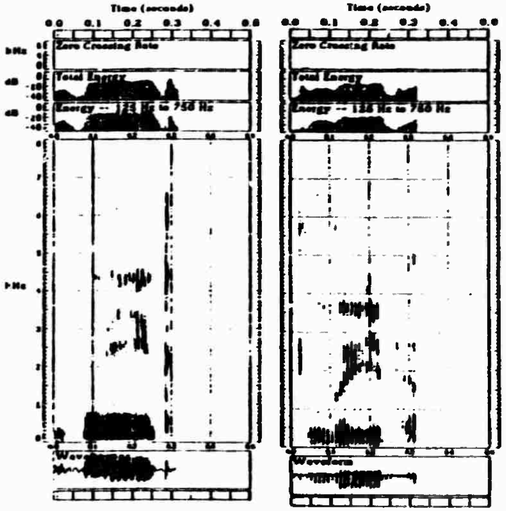


Figure 5.1: Spectrograms of the words *need*, and *lead*

In order to simplify the nasal recognition problem, nasal consonant boundaries were detected manually in this research. Clearly, it would be worthwhile establishing some automatic procedure for detecting nasal consonant boundaries. A preliminary feasibility study examined the performance of a boundary detection algorithm based on locating points of maximum spectral change in the speech waveform. A similar procedure was used successfully in the past by Mermelstein [48]. An evaluation of all nasal vowel sequences in the database, the results of which are presented in figure 5.2, showed that nasal consonant boundaries can be located within 20 msec of a manually assigned boundary over 95% of the time.

The experiments in nasalized vowel detection indicated that it is possible to use

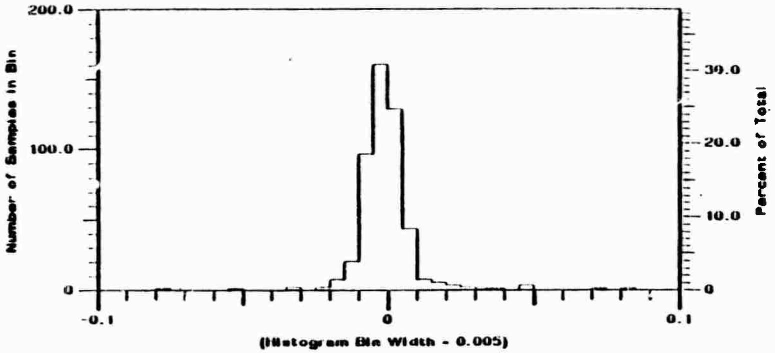


Figure 5.2: A Histogram of Boundary Detection Error

This figure contains a histogram of the errors of the automatic nasal consonant boundary detection algorithm. The error is calculated by taking the time difference between the hypothesized boundary, and a manually assigned boundary. Values are in seconds.

information in the vowel to predict the presence of a nasal consonant. Thus, it would be worthwhile to establish if this information actually improves nasal detection systems. Finally, it would be interesting to examine the usefulness of speaker adaptation in the nasalized vowel detection system, since vowel nasalization was found to be strongly speaker dependent.

Bibliography

[1] Ali, A., Gallagher, T., Goldstein, J., Daniloff, R., "Perception of Coarticulated Nasality, *Journal of the Acoustical Society of America*, Vol. 49, No. 2, pp. 538-540, 1971.

[2] Bickley, C.A., "Acoustic Analysis and Perception of Breathy Vowels", *Working Papers, Speech Communication Group*, Research Laboratory of Electronics, MIT, Vol. 1, pp. 71-82, 1982.

[3] Blomberg, M., Carlson, R., Elenius, K., Granstrom, B., "Auditory Models in Isolated Word Recognition, *Proceedings ICASSP 84*, San Diego, CA, pp. 17.9.1-17.9.4, 1984.

[4] Blumstein, S.E., Stevens, K.N., "Acoustic Invariance in Speech Production: Evidence from Measurements of the Spectral Characteristics of Stop Consonants, *Journal of the Acoustical Society of America*, Vol. 66, No. 4, pp. 1001-1017, 1979.

[5] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., *Classification and Regression Trees*, Wadsworth International Group, Belmont, California, 1984.

[6] Bush, M.A., Kopec, G.E., Zue, V.W., "Selecting Acoustic Features for Stop Consonant Identification, *Proceedings of ICASSP 83*, Boston, MA, 1983, pp. 742-745.

[7] Cole, R.A., Editor *Perception and Production of Fluent Speech*, Lawrence-Erlbaum Ass., Hillsdale, New Jersey, 1980.

[8] Dautrich, B.A., Rabiner, L.R., Martin, T.B., "On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition, *IEEE Transactions ASSP*, Vol. 31, No. 4, pp. 793-806, 1983.

[9] De Mori, R., Gubrynowicz, R., Laface, P., "Inference of a Knowledge Source for the Recognition of Nasals in Continuous Speech, *IEEE Transactions ASSP*, Vol 5, pp. 538-549, 1979.

[10] Dickson, D.R., "Acoustic Study of Nasality, *Journal of Speech and Hearing Research*, Vol. 5, No. 2, pp. 103-111, 1962.

[11] Dixson, N.R., Silverman, H.F., "A General Language Operated Decision Implementation System (GLODIS): Its Application to Continuous-Speech Segmentation, *IEEE Transactions ASSP*, Vol. 24, pp. 137-162, 1976.

[12] Fairbanks, G., House, A.S., Stevens, A.L., "An Experimental Study of Vowel Intensities, *Journal of the Acoustical Society of America*, Vol. 22, No. 4, pp. 457-459.

[13] Fant, G., *Acoustic Theory of Speech Production*, Mouton and Co., 's-Gravenbage, Netherlands, 1960.

[14] Flanagan, J.L., *Speech Analysis, Synthesis and Perception*, Springer-Verlag, New York, 1972.

[15] Fujimura, O., "Spectra of Nasalized Vowels, *Research Laboratory of Electronics, MIT Quarterly Report*, No. 58., pp. 214-218, 1960.

[16] Fujimura, O., "Analysis of Nasal Consonants, *Journal of the Acoustical Society of America*, Vol. 34, No. 12, pp. 1865-1875, 1962.

[17] Fujimura, O., "Formant-Antiformant Stucture of Nasal Murmers, *Proceedings of the Speech Communication Seminar*, Vol 1, Stockholm: Royal Institute of Technology, Speech Transmission Laboratory, pp. 1-9, 1962.

[18] Fujimura, O., Lindqvist, J., "Sweep-Tone Measurements of the Vocal Tract Characteristics, *Journal of the Acoustical Society of America*, Vol. 49, No. 2, pp. 541-558, 1971.

[19] Gillmann, R.A., "Automatic Recognition of Nasal Phonemes, *Proceedings IEEE Symposium on Speech Recognition*, Pittsburgh, PA, 1974, pp. 74-79.

[20] Hattori, S., Yamamoto, K., Fujimura, O., "Nasalization of Vowels in Relation to Nasals, *Journal of the Acoustical Society of America*, Vol. 30, No. 4, pp. 267-274, 1958.

[21] Hawkins, S., Stevens, K.N., "A cross-language study of the perception of nasal vowels, Paper presented at the 105th meeting of the Acoustical Society of America, Cincinatti, Ohio, 1983.

[22] Hess, W.J., "A Pitch-Synchronous Digital Feature Extraction System for Phonemic Recognition of Speecb, *IEEE Transactions ASSP*, Vol. 24, pp 14-25, 1976.

[23] House, A.S., Stevens, K.N., "Analog Studies of the Nasalization of Vowels, *Journal of Speech and Hearing Disorders*, Vol. 22, No. 2, pp. 218-232, 1956.

[24] House, A.S., "Analog Studies of Nasal Consonants, *Journal of Speech and Hearing Disorders*, Vol. 22, pp. 190 204, 1957.

[25] Hyde, S.R., "Automatic Speech Recognition: A Critical Survey and Discussion of the Literature, *Human Communication: A Unified View*, edited by E.E. David and P.B. Denes, McGraw-Hill, New York, 1972.

[26] Jakobson, R., Fant, G., Halle, M., *Preliminaries to Speech Analysis*, MIT Press, Cambridge, Mass, 1963.

[27] Kawasaki, H., "The perceived nasality of vowels with gradual attenuation of adjacent nasal consonants, Paper presented at joint meeting of the Acoustical Society of America and Acoustical Society of Japan, Honolulu, HI, 1978.

[28] Klatt, D.H., "Review of the ARPA Speech Understanding Project, *Journal of the Acoustical Society of America*, Vol. 62, No. 6, pp. 1345-1366, 1977.

[29] Klatt, D.H., "Speech Perception: a Model of Acoustic-Phonetic Processing and Lexical Access, *Journal of Phonetics*, Vol. 7, pp. 279-312, 1979.

[30] Kopec, G.E., "Voiceless Stop Consonant Identification Using LPC Spectra, *Proceedings of ICASSP 84*, San Diego, CA, 1984.

[31] Kurowski, K., Blumstein, S.E., "Perceptual Integration of the Murmur and Formant Transitions for Place of Articulation in Nasal Consonants, *Journal of the Acoustical Society of America*, Vol. 76, No. 2, pp. 383-390, 1984.

[32] Labov, W., Yaeger, M., Steiner, R., "A Quantative Study of Sound Change in Progress", Report on National Science Foundation Contract, NSF-65-3287, University of Pennsylvania, 1972.

[33] Larkey, L.S., Wald, J., Strange, W., "Perception of synthetic nasal consonants in initial and final syllable position, *Perception and Psychophysics*, Vol. 23, No. 4, pp. 299-312, 1978.

[34] Lea, W.A., Editor *Trends is Speech Recognition*, Prentice Hall Inc., Englewood Cliffs, New Jersey, 1980.

[35] Leung, H.C., Zue, V.W., "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech, *Proceedings of ICASSP 84*, San Diego, CA, 1984.

[36] Liberman, A.M., Delattre, P., Cooper, F.S., Gerstman, L.J., "The Role of Consonant-Vowel Transitions in the Perception of the Stop and Nasal Consonants, *Psychological Monographs*, Vol. 68, No. 8, pp. 1-13, 1954.

[37] Lindqvist, J., Sundberg, J., (1972), "Acoustic Properties of the Nasal Tract, *Speech Transmission Laboratory Quarterly Progress Status Report*, No. 1, Royal Institute of Technology, Stockholm, pp. 13-17, 1972.

[38] Lintz, L.B., Sherman, D., "Phonetic Elements and Perception of Nasality, *Journal of Speech and Hearing Research*, Vol. 4, No. 4, pp. 381-396, 1961.

[39] Maeda, S., "The Role of the Sinus Cavities in the Production of Nasal Vowels, *Proceedings of ICASSP 82*, Paris, France, 1982, pp. 911-914.

[40] Maeda, S., "Acoustic Correlates of Vowel Nasalization: a Simulation Study, Paper presented at the 104th meeting of Acoust. Soc. Amer., Orlando, FL, 1982.

[41] Makhoul, J.I., Wolf, J.J., "Linear Prediction and the Spectral Analysis of Speech, Bolt, Beranek and Newman Report No. 2304, 1972.

[42] Makhoul, J.I., "Linear Prediction: A Tutorial Review, *Proc. IEEE*, Vol. 63, pp. 561-580, April 1975.

[43] Malécot, A., "Acoustic Cues for Nasal Consonants: An Experimental Study Involving a Tape-Splicing Technique, *Language*, Vol. 32, pp. 274-284, 1956.

[44] Malécot, A., "Vowel Nasality as a Distinctive Feature in American English, *Language*, Vol. 36, No 2, pp. 222-229, 1960.

[45] Markel, J.D., Gray Jr., A.H., *Linear Prediction of Speech*, Springer-Verlag, New-York, 1976.

[46] Mártony, J., "The role of formant amplitudes in synthesis of nasal consonants, *Speech Transmission Laboratory Quarterly Progress Status Report*, No. 3, Royal Institute of Technology, Stockholm, pp. 28-31, 1964.

[47] Mathews, M.V., Miller, J.E., David Jr., E.E., "Pitch Synchronous Analysis of Voiced Sounds, *Journal of the Acoustical Society of America*, Vol. 33, No. 2, pp. 179-186, 1961.

[48] Mermelstein, P., "On Detecting Nasals in Continuous Speech, *Journal of the Acoustical Society of America*, Vol. 61, No. 2, pp. 581-587, 1977.

[49] Miller, G.A., Nicely, P.E., "An Analysis of Perceptual Confusions Among Some English Consonants, *Journal of the Acoustical Society of America*, Vol. 27, No. 2, pp. 338-352, 1955.

[50] Nakata, K., "Synthesis and Perception of Nasal Consonants, *Journal of the Acoustical Society America*, Vol. 31, No. 6, pp. 661-666, 1959.

[51] Nguyen, D.T., Guern, B., "Effects of Nasal Coupling on the Vowels, Paper presented at the 99th meeting of the Acoustical Society of America, Atlanta, GA, 1980.

[52] Noll, A.M., "Cepstrum Pitch Determination, *Journal of the Acoustical Society of America*, Vol. 41, No. 2, pp. 293-309, 1967.

[53] Nord, L., "Experiments with Nasal Synthesis, *Speech Transmission Laboratory Quarterly Progress Status Report*, No. 2-3, Royal Institute of Technology, Stockholm, pp. 14-19, 1976.

[54] Nord, L., "Perceptual Experiments with Nasals, *Speech Transmission Laboratory Quarterly Progress Status Report*, No. 2-3, Royal Institute of Technology, Stockholm, pp. 5-8, 1976.

[55] Oppenheim, A.V., "Speech analysis-synthesis system based on homomorphic filtering, *Journal of the Acoustical Society of America*, Vol. 45, No. 2, pp. 458-465, 1969.

[56] Oppenheim A.V., Schafer, R.W., *Digital Signal Processing*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1975.

[57] Oppenheim A.V., *Applications of Digital Signal Processing*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1978.

[58] Pinson, E.N., "Pitch Synchronous Time-Domain Estimation of Formant Frequencies and Bandwidths, *Journal of the Acoustical Society of America*, Vol. 35, No. 8, pp. 1264-1273, 1963.

[59] Peterson, G.E., Barney, H.L., "Control Methods Used in a Study of the Vowels, *Journal of the Acoustical Society of America*, Vol. 24, No. 2, pp. 175-185, 1952.

[60] Pols, L.C., Tromp, H.R.C., Plomp, R., "Frequency Analysis of Dutch Vowels from 50 Male Speakers, *Journal of the Acoustical Society of America*, Vol. 53, No. 4, pp. 1093-1101, 1973.

[61] Rabiner, L.R., Schafer, R.W., *Digital Processing of Speech Signals*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1978.

[62] Raphael, L., Dormann, M., Freeman, F., "Vowel and Nasal Duration as Cues to Voicing in Word-Final Stop Consonants: Spectrographic and Perceptual Studies, *Journal of Speech and Hearing Research* Vol. 18, pp. 839-400, 1975.

[63] Recasens, D., "Place Cues for Nasal Consonants with special reference to Catalan, *Journal of the Acoustical Society of America*, Vol. 73, No. 4, pp. 1346-1353, 1983.

[64] Repp, B.H., "Perception of the [m]-[n] Distinction: Insights from Four Converging Procedures, Paper presented at 108th meeting of the Acoustical Society of America, Minneapolis, Minnesota, 1984.

[65] Rosenberg, A.E., "Effect of Glottal Pulse Shape on the Quality of Natural Vowels, *Journal of the Acoustical Society of America*, Vol. 49, No. 2, pp. 583-590, 1971.

[66] Schafer, R.W., Rabiner, L.R., "System for automatic formant analysis of voiced speech, *Journal of the Acoustical Society of America*, Vol. 47, No. 2, pp. 634-648, 1970.

[67] Searle, C.L., Jacobson, J.Z., Rayment, S.G., "Stop consonant discrimination based on human audition, *Journal of the Acoustical Society of America*, Vol. 65, No. 3, pp. 799-809, 1979.

[68] Seneff, S., Klatt, D.H., Zue, V.W., "Design considerations for optimizing the intelligiblity of DFT-based, pitched-excited, critical-band spectrum speech analysis/resynthesis system, *Speech Communication Group Working Papers*, No. 1, pp. 31-46, 1982.

[69] Shipman, D.W., "SpireX: Statistical Analysis in the Spire Acoustic-Phonetic Workstation, Proceedings of ICASSP 83, Boston, MA, pp. 1360-1363, 1983.

[70] Stevens, K.N., Klatt, M., "Study of Acoustic Properties of Speech Sounds, Bolt, Beranek and Newman Report No. 1669, 1968.

[71] Stevens, K.N., "Study of Acoustic Properties of Speech Sounds II, and Some Remarks on the Use of Acoustic Data in Schemes for Machine Recognition of Speech, Bolt, Beranek and Newman Report No. 1871, 1969.

[72] Su, L.S., Li, K.P., Fu, K.S., "Identification of speakers by the use of nasal coarticulation, *Journal of the Acoustical Society of America*, Vol. 56, No. 6, pp. 1876-1882, 1974.

[73] Weinstein, C.J., McCandless, S.S., Mondshein, L.F., Zue, V.W., "A System for Acoustic-Phonetic Analysis of Continuous Speech, *IEEE Transactions ASSP*, Vol. 23, pp. 54-67, 1975.

[74] White, G.M., Neely, R.B., "Speech recognition experiments with linear prediction, handpass filtering, and dynamic programming, *IEEE Transactions ASSP*, Vol. 24, No. 2, pp. 183-188, 1976.

[75] Wright, J., "The Behavior of Nasalized Vowels in the Perceptual Vowel Space, *Report of the Phonology Laboratory*, No. 5, Berkely, CA, 1980.

[76] Zue, V.W., "Acoustic Characteristics of Stop Consonants: A Controlled Study, Sc.D. Thesis, M.I.T., 1976.

[77] Zue, V.W., Laferriere, M., "Acoustic Study of Medial /t,d/ in American English, *Journal of the Acoustical Society of America*, Vol. 66, No. 4, pp. 1039-1050, 1979.

[78] Zue, V.W., "Acousic-Phonetic Knowledge Representation: Implications from Spectrogram Reading Experiments, *Proceedings of the 1981 NATO Advanced Summer Institute on Automatic Speech Analysis and Recognition*, P. Reidel Publishing Co., 1981.

[79] Zue, V.W., Sia, E.B., "Nasal Articulation in Homorganic Clusters in American English, Paper presented at 102nd meeting of the Acoustical Society of America, Miami, FL, 1981.

[80] Zwicker, E., Terhardt, E., Paulus, E., "Automatic speech recognition using psychoacoustic models, *Journal of the Acoustical Society of America*, Vol. 65, No. 2, pp. 487-498, 1979.

Appendix A

Corpus Words

The corpus has a total of 203 different words containing nasal consonants in various positions, both as a single consonant and as part of a cluster. Care was taken to include words that formed minimal pairs, as well as words with acoustic characteristics that are similar to nasals. The following tables section the corpus words into their basic phonetic environment.

Table A.1: Consonant Nasal Clusters

m	n	w/o nasal	across syllable		similar words				
smack	snack	sack			slack	nack	mack	lack	back
smoke	snowed	sewed	gismo		slowed	note	low	moat	dote
smock	snot	sought			mod	slot	lot	not	swat
smitten	suit	sit	ethnic	parsnip	slit	lit	knit	mitt	
film	kiln	kill							
dorm	corn	whorl							

Table A.2: Nasal Stop Consonant Clusters

with stop		w/o nasal		w/o stop	across syllable		similar words		
camp		cap	cab	cam	camper	campbell	can		
sump		sup	sub	some	somebody		sun	sung	
font	fond	fought		fawn	fondest		fall	fault	
bent	bend	bet	bed	ben	bending	sentry	bell	belt	
pant	panned	pat	pad	pan	panter	pander	pal	pam	
sink		sick		sing	sinking	sing	sill	sin	silk
sunk		suck		sung	sunken	hunger	sun	some	sulk

Table A.3: Syllable Initial Nasals

m	n	similar words			across syllable		
made	nape	bade	tape		helpmate	cognate	
mitt	nip	bit	dip	lip	abnegate	admit	picnic
meat	need	beat	deed	lead	voltmeter	technique	
mack	nack	back	tack	lack	enigma		
moat	note	boat	dote	vote	utmost	ignore	
mutt	nut	but	dud		chipmunk	pignut	

Table A.4: Syllable Final Nasals

m	n	ŋ	similar words			across syllable	
cam	can		cab	cad		campbell	
dim	din	ding	dip	did	dill	skimpy	
some	sun	sung	sub	sud		sunken	somebody
comb	bone		cope	bode	bowl	lonely	homely

Table A.5: Nasal Fricative Clusters

w/o fricative	with fricative		across syllable	similar words	
warm	warms	warmth		worn	
	triumph		triumphant		
limb	lymph				
won	once	ones			
pin	pinch	pins	pinching	pill	pills
strain	strange	strains	stranger		
string	strength	strings			

Table A.6: Intervocalic Nasals

m	n	ŋ	similar words	
simmer	sinner	singer	tiller	critter
hammer	banner	banger	matter	
rummy	runny		ruddy	sully
comic	conic		polish	
demise	denies		relies	devise
hammock	bannock		baddock	bavock

Table A.7: Syllabic Nasals

m	n	similar word
bottom	button	bottle
totem	oaten	total

Table A.8: Miscellaneous

chimney
inmate
hangman
omnibus
damnation
dalmation
arsenic
decimal
animal
flannel

Appendix B

Phonetic Transcription Alignment Procedures

In order to be able to analyze utterances with the SpireX statistical package, time aligned phonetic transcriptions were required. This appendix describes the procedure for time alignment, and summarizes the rules used.

In this research, the phonetic transcriptions were aligned manually to the waveform using the Spire facility available on MIT Lisp Machine work stations. A typical transcription layout, illustrated in figure B.1, contains:

1. the orthographic, and phonetic transcriptions,
2. a menu of possible phonetic symbols,
3. a broad-band spectrogram of the utterance,
4. one compressed, and one expanded view of the speech waveform and,
5. a short-time spectral slice, computed with a 6.6 msec hamming window, at the position of a time cursor.

By positioning a cursor and a marker, a segment region may be established. As shown in figure B.1, this segment region is indicated by two vertical lines on the

spectrogram, or speech waveforms. A phonetic symbol is associated with this time segment by selecting an element from the symbol table.

The time alignment process usually proceeds from left to right. In a typical alignment operation, the spectrogram is used to position the time cursor near the next segment boundary. The exact position of the boundary is determined by observation of the expanded speech waveform, the short-time spectra, and when necessary, by listening to the speech segment.

The boundary between a nasal consonant and an adjacent sonorant is not difficult to establish since it is denoted by sudden spectral and intensity changes. This is reflected by sharp changes in the spectrogram, as shown in figure B.2 for the word *simmer*. Note that the periodic waveform changes its shape noticeably on either side of the boundary. For these types of transitions the boundary was set at the point of maximum spectral change.

The boundary between a nasal consonant and a voiceless obstruent, or period of silence, was set at the onset (or offset) of voicing in the waveform. Figure B.3 illustrates the case for the fricative nasal cluster in the word *smack*. Here, the boundary was set at the onset of voicing of the nasal consonant. The period of epenthetic silence in between the fricative and the nasal consonant is caused by an asynchrony mistiming of the movements of the articulators, and is common in all fricative nasal clusters.

The boundary between a nasal consonant and a voiced obstruent was determined by an onset of some other acoustic characteristic. Figure B.4 illustrates the case of the word *warms*, where the boundary is set at the onset of frication. Figure B.5 illustrates the case of the word *bending*, where the boundary was determined by the lack of energy immediately above the low frequency resonance (relative to the rest of the murmur, which was labeled as the nasal consonant). Note that this difference is quite subtle.

The most difficult boundaries to establish were between nasal consonants and voice bars, as shown in the previous figure, or between two different nasal consonants, as illustrated in figure B.6, for the word *inmate*. In these cases, the boundary was determined by observing subtle chang... ı the short-time spectra, and by listening to the individual regions in the speech waveform.

In many intervocalic environments, the consonant /n/ was produced as a nasal flap, as shown in figure B.7, for the word *bannock*. For transcription purposes, any /n/, in an intervocalic environment, which was less than two pitch periods long, was labeled a nasal flap.

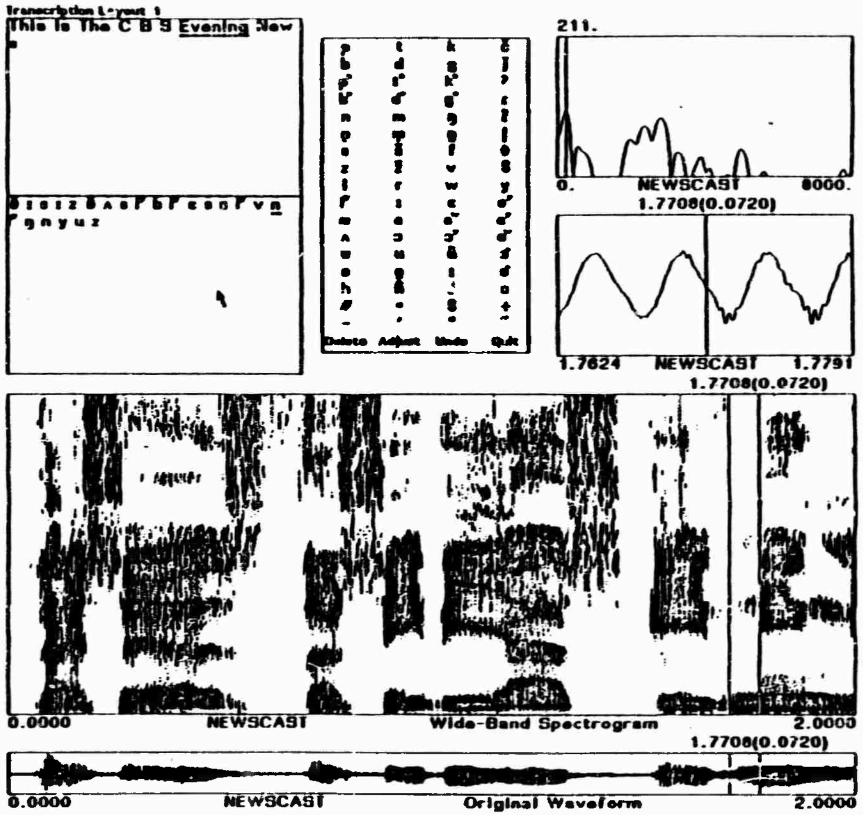


Figure B.1: The Spire Phonetic Transcription Layout

This figure contains a typical transcription layout of the Spire facility. The layout contains, counter-clockwise from the upper left, the orthographic transcription, the phonetic transcription, a broad-band spectrogram, the speech waveform, an expanded speech waveform, and a short-time spectra computed at the time of the cursor. A phonetic symbol table is located in the middle of the layout. The cursor is indicated by a vertical line in the spectrogram, or speech waveform, and is controlled by the mouse.

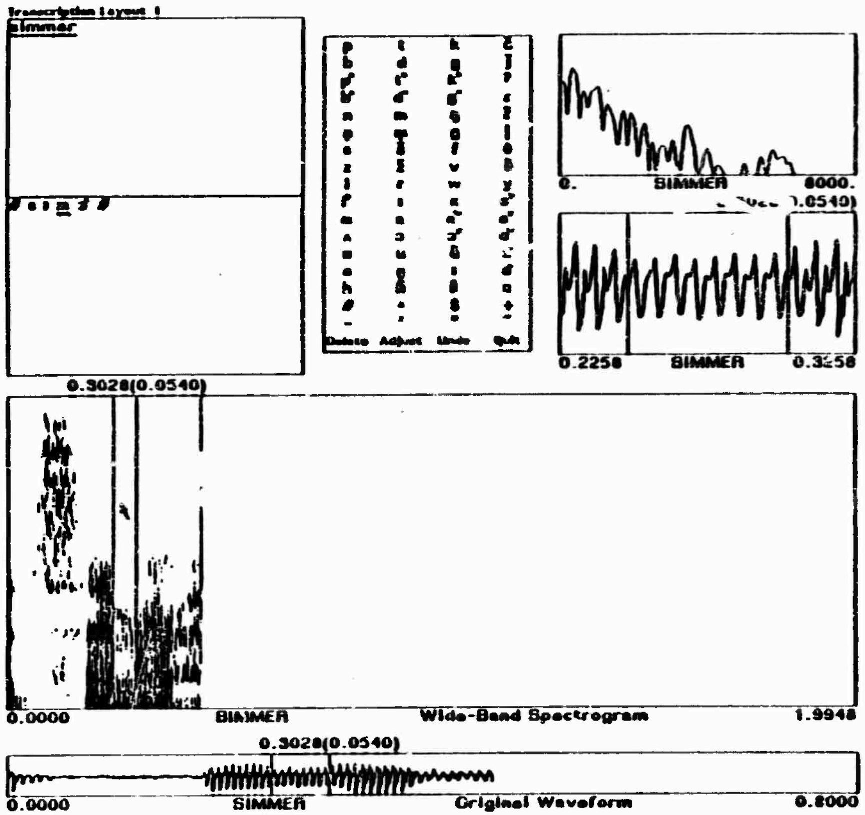


Figure B.2: The Traoscription of the word *simmer*

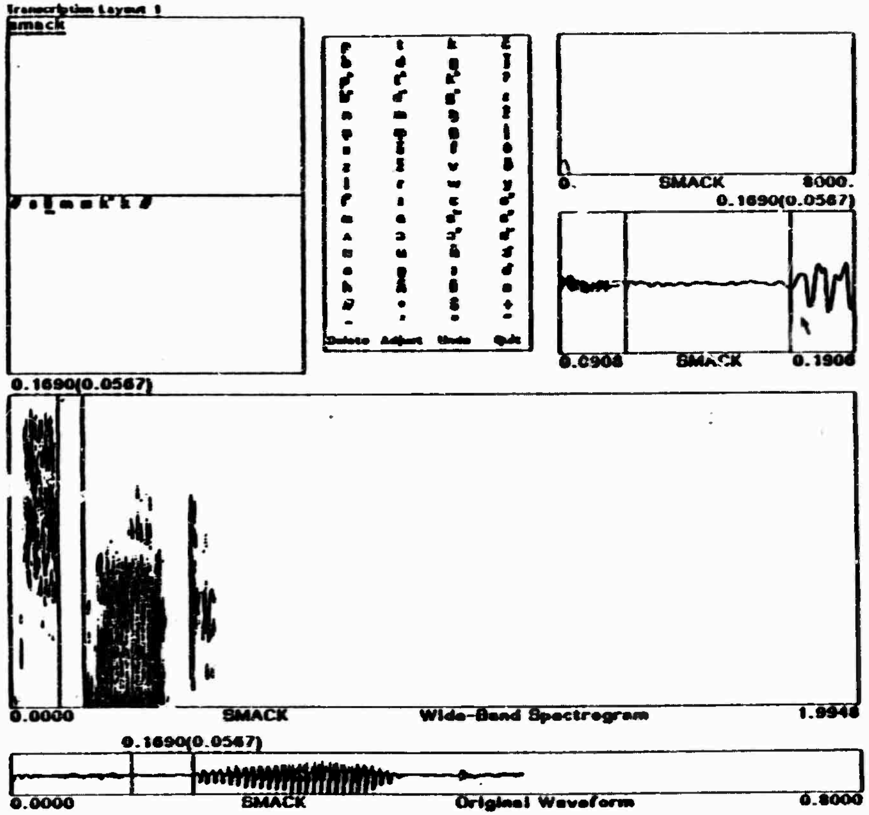


Figure B.3: The Transcription of the word *smack*

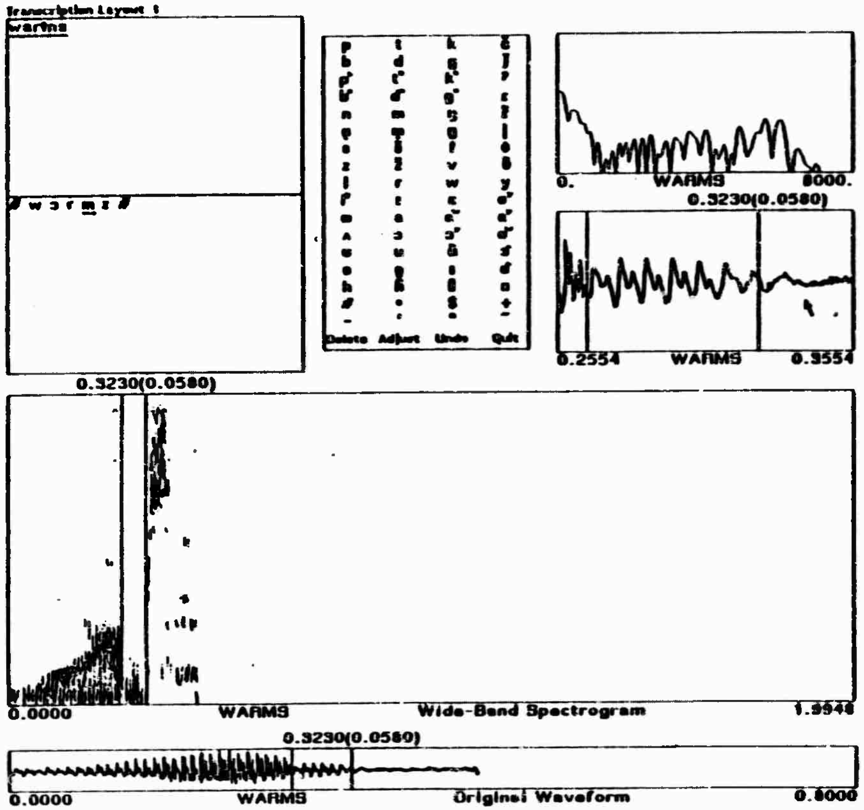


Figure B.4: The Transcription of the word *warms*

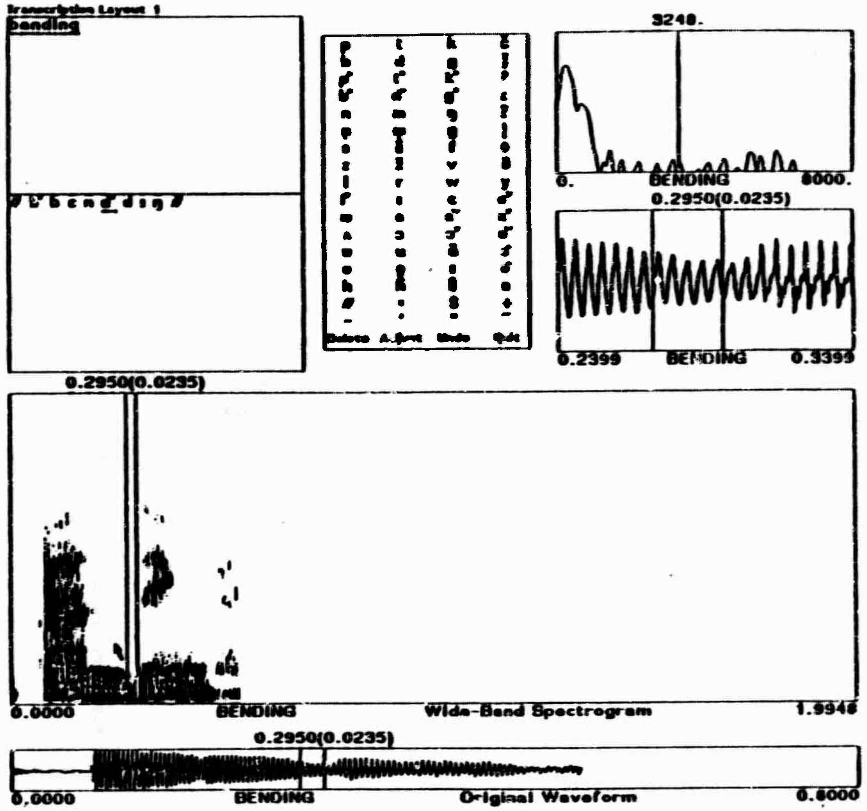


Figure B.5: The Transcription of the word *bending*

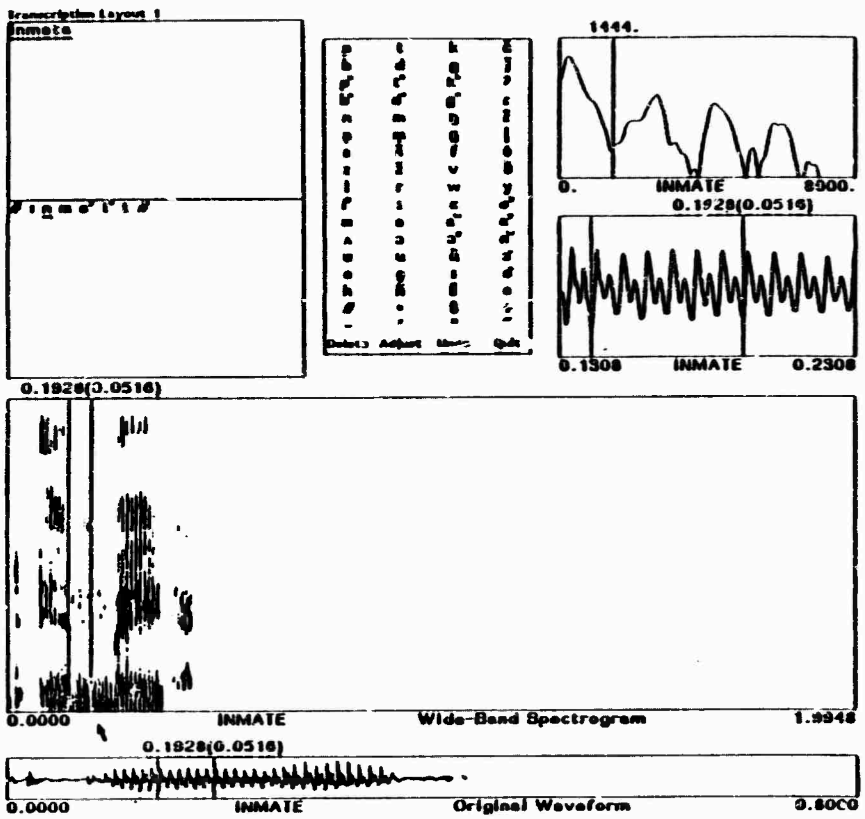


Figure B.6: The Transcription of the word *inmate*

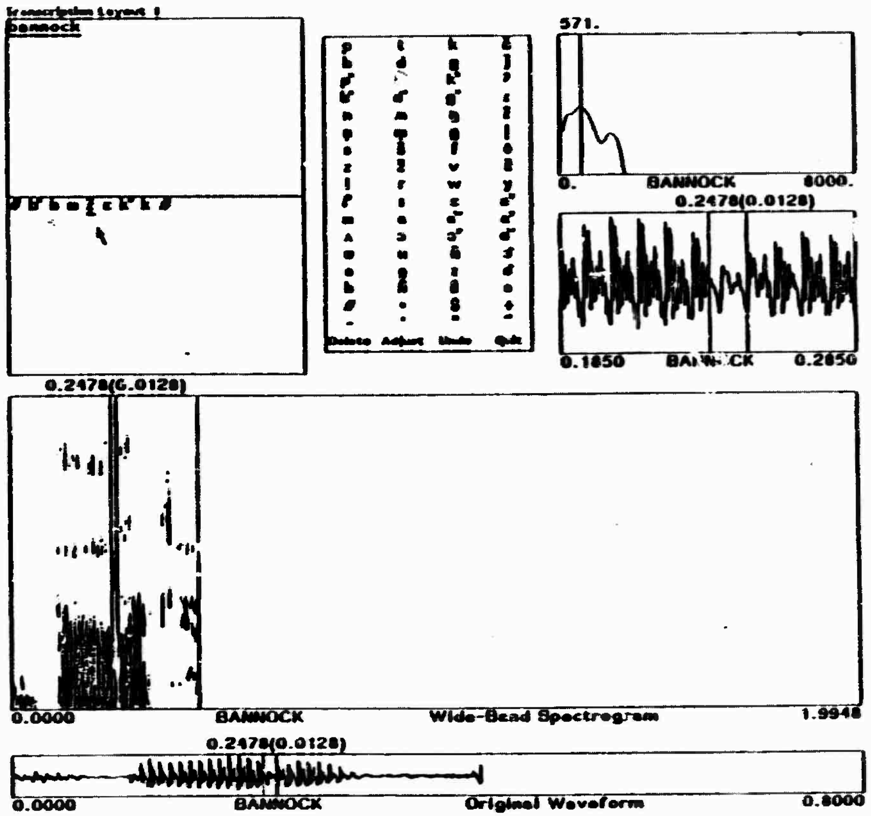


Figure B.7: The Transcription of the word *bannock*

the short-time Fourier transform can be shown to be equivalent to the discrete Fourier transform (DFT) of the windowed sequence and thus can be computed using the fast Fourier transform (FFT) algorithm [56]. This points out the fundamental similarities between a filter bank and the DFT.

Since the shape of the filter windows has a substantial effect on the output of the short-time Fourier transform, it is important to consider them carefully. The remainder of this section will focus on this issue, given that we have decided to create a filter bank for spectral analysis purposes. The type of filter bank necessary for a statistical analysis will also be discussed.

Although a number of different filter bank structures have been proposed for speech analysis, there is no simple guideline for choosing an optimal filter bank for a particular application. There are many different variables to determine including: the type of filter (IIR or FIR [61]); the filter spacing (uniform or nonuniform; overlapping or nonoverlapping); the number of filters and the filter frequency responses. One constraint which limits the range of some of these parameters is an intelligibilty requirement. The filter bank output should retain enough information of the original speech signal that it can be correctly perceived by human listeners. Thus one could evaluate the relative merit of different filter banks by performing a series of perceptual tests [68]. A similar form of judgment could be made by analyzing vocoder performances [14]. Another possible procedure to judge the merit of different filter banks would be to investigate their relative success at the front end of a standard speech recognition system [3], [8], [74]. Using this latter method, Dautrich et al. have reported a number of interesting points concerning filter banks:

- filter bank performance deteriorates for too few filters due to poor resolution.
- filter bank performance deteriorates for too many nonoverlapping filters since their frequency responses become so narrow that some end up measuring noise between pitch harmonics.

- successful filter banks have an essentially flat overall frequency response without sharp peaks or valleys.
- the best performance of non-uniform filter banks is obtained for filters spaced along a critical band frequency scale as opposed to octave bands, $\frac{1}{3}$ octave bands, or arbitrary spacing.

Of course, it is possible to design a filter bank solely on psychophysical data. The motivation for this is that a spectral analysis would then emphasize the things which are known to be perceptually important and demphasize those which are not. For instance, a small change in the frequency of a higher formant should not be as important as the same change in the first formant because the just-noticable difference (JND) for a formant frequency increases with frequency [29]. Several attempts have been made to design filter banks with these considerations in mind [67],[80].

What kind of filter bank could be used for a statistical analysis of speech sounds? Optimally, the exact filter bank shape should not be critical to the success of an analysis experiment. Global spectral features should be able to be established by the data independent of any particular filter bank shape. Specific filter bank details could be dictated somewhat by the particular kind of spectral analysis being performed. For an analysis of the steady state portion of nasal murmurs or nasalized vowels for instance, a long filter impulse response or time window could be used since there are no sharp temporal changes in these regions. The main advantage of using a long filter window is that one can obtain good spectral stability independent of the window position relative to the pitch period as illustrated in figures C.1 – C.3 for a synthetic steady-state vowel with a fundamental frequency of around 100 Hz.

In figure C.1 we see that two hamming windows (with 7 ms and 25 ms duration) have been centered at the beginning of a pitch period. Both of their corresponding DFT's show a good spectral representation of the vocal tract. Note that the pitch harmonics are visible in the DFT of the 25 msec hamming window because of the tradeoff between time and frequency resolution. Figure C.2 illustrates the same conditions except that the windows have been centered at the tail end of the preceeding pitch period. As one would expect, the DFT of the shorter hamming window yields a very poor spectral representation of the resonances of vocal tract. At this point in the pitch period, the glottal folds are open so that the resonances of the vocal tract are severely damped. Figure C.2 also shows that the longer hamming window is able to extract a reliable vocal tract shape since it overlaps multiple pitch periods. Figure C.3 illustrates the effect of centering the windows in the middle of a pitch period. Clearly a longer window length will yield a more stable spectral response.[2]

The penalty for spectral stability is reduced temporal resolution as is illustrated in figures C.4 – C.6 for another synthetic vowel. In this example the first and third resonances are held fixed at 450 and 2450 Hs respectively while the second resonance is changed from 950 to 1950 Hs within 10 msec. In figure C.4 the windows are centered at the start of the last pitch period before the start of the transition period. Note that the DFT of the longer hamming window reflects the fact that the window overlaps multiple pitch periods with different vocal tract characteristics since the second resonance range is smeared in the spectral domain. This holds true for figures C.5 and C.6 as well. Notice that in each case the shorter hamming window isolates a single pitch period which produces a superior spectral shape.

Clearly if we knew the exact location of the pitch period, it would be possible to center a window over the important part of the pitch period to produce an excellent estimate of the vocal tract spectral shape. The duration of the window could be chosen to be short enough so that there would be no pitch information present in the spectral domain resulting in a smooth spectral shape. Further, the spectra would be stable with time yet accurately reflect changes in the vocal tract.

[2] A 25 msec window starts to show instability if the pitch period is much below 12.5 msec (80 Hz).

Unfortunately, automatic pitch-synchronous analysis is difficult to perform reliably. Most analysis procedures locate the pitch period boundaries either manually or at most semi-automatically [47], [58], [65]. In any event, automatic pitch-synchronous analysis is beyond the scope of this thesis. As was illustrated in the previous figures, a long window is probably the best choice for asynchronous analysis of steady state sounds.

For spectral analysis of nasal consonants and nasalised vowels it seems reasonable to use a long window which would provide good spectral resolution. Temporal resolution is not a critical factor here because the analysis will take place in a relatively stationary environment.

As a first pass at analysis, a uniformly spaced filter bank was used. A hamming window was chosen because of its superior spectral properties.[3] For spectral stability for most pitch frequencies a long duration window (25 msec) was used.

From previous figures it is clear that with decreased time resolution one obtains superior spectral resolution. However this is detrimental to any study of spectral shapes since one does not want pitch information in the spectral estimate of the vocal tract. Thus some form of spectral smoothing is necessary. One solution to this problem would be to smooth the cepstrum of the speech signal.[4] Cepstral smoothing or homomorphic filtering has been successfully used for many speech processing applications where smoothed spectra are necessary [52], [55], [66]. Figures C.7 – C.9 show a plot of the cepstrum, the cepstrum window and the unsmoothed and smoothed FFT spectra for various pitch values, ranging from 120 Hs to 440 Hz, on a synthetic vowel. For the purposes of this analysis the smoothing window implemented was 3 msec long (flat for 1.5 msec and tapered with a raised cosine for 1.5 msec). This window was found to produce acceptably smooth spectra for pitch frequencies below 300 Hz. This is acceptable for most

[3] This data window is attractive because the side lobes of its Fourier transform remain more that 40 dB down at all frequencies [56].

[4] The cepstrum is defined as the Inverse Fourier transform of the log magnitude spectrum [55].

male and female speakers but not for pitch frequencies of some children. The problem is avoided since the speech of children is not is not analysed in this study.

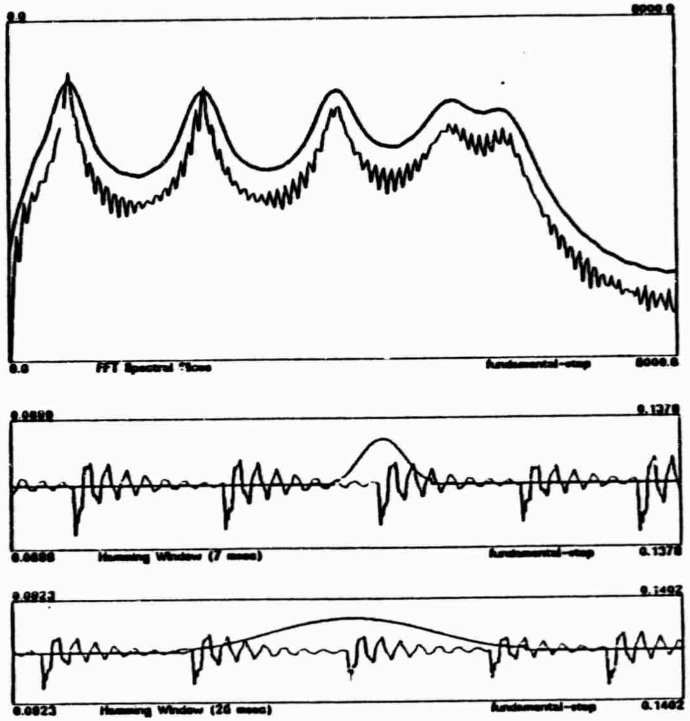


Figure C.1: Hamming Windows Centered at Start of Pitch Pulse

The top display contains DFT spectra for two different duration hamming windows (7 msec and 25 msec). The thick line corresponds to the shorter hamming window. The other displays contain the hamming windows and the original speech waveform.

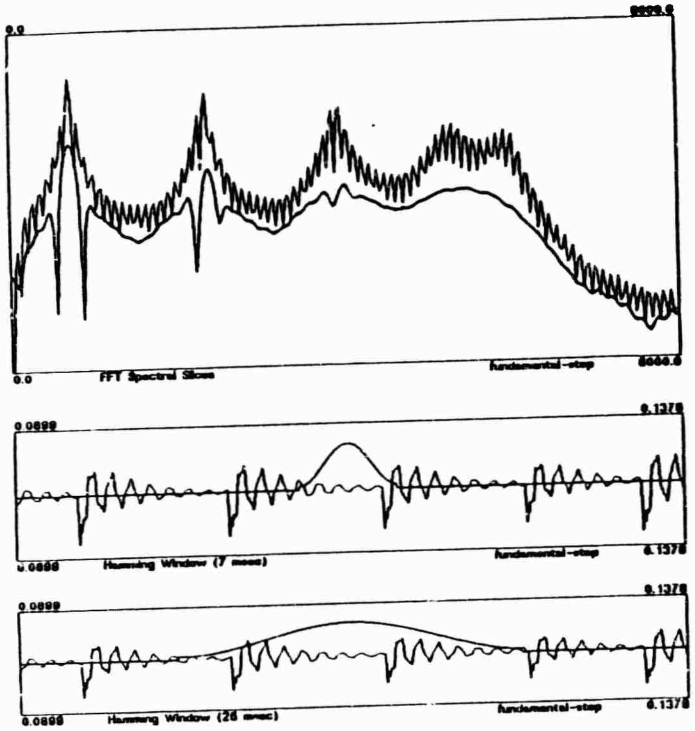


Figure C.2: Hamming Windows Centered at End of Pitch Pulse

The top display contains DFT spectra for two different duration hamming windows (7 msec and 25 msec). The thick line corresponds to the shorter hamming window. The other displays contain the hamming windows and the original speech waveform.

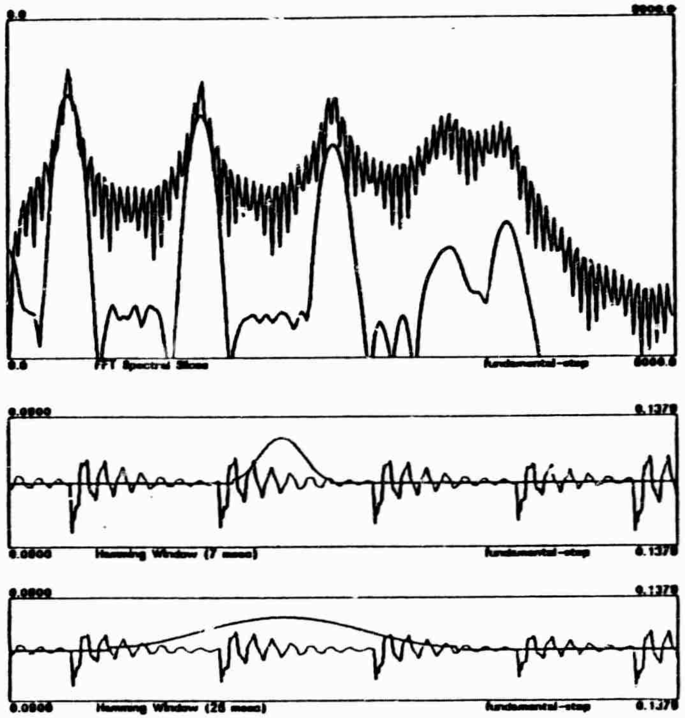


Figure C.3: Hamming Windows Centered at Middle of Pitch Pulse

The top display contains DFT spectra for twq different duration hamming windows (7 msec and 25 msec). The thick line corresponds to the shorter hamming window. The other displays contain the hamming windows and the original speech waveform.

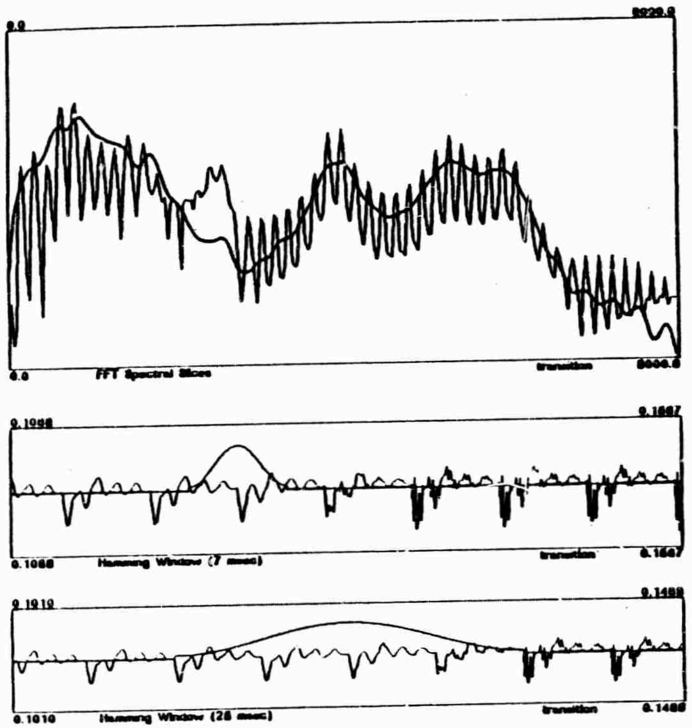


Figure C.4: Hamming Windows Centered at Start of Formant Transition

The top display contains DFT spectra for two different duration hamming windows (7 msec and 25 msec). The thick line corresponds to the shorter hamming window. The other displays contain the hamming windows and the original speech waveform.

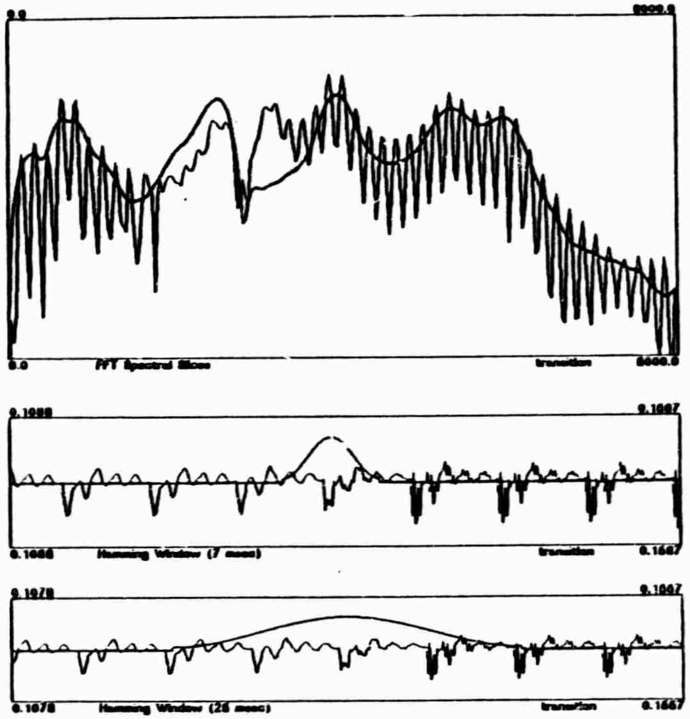


Figure C.5: Hamming Windows Centered at Middle of Formant Transition

The top display contains DFT spectra for two different duration hamming windows (7 msec and 25 msec). The thick line corresponds to the shorter hamming window. The other displays contain the hamming windows and the original speech waveform.

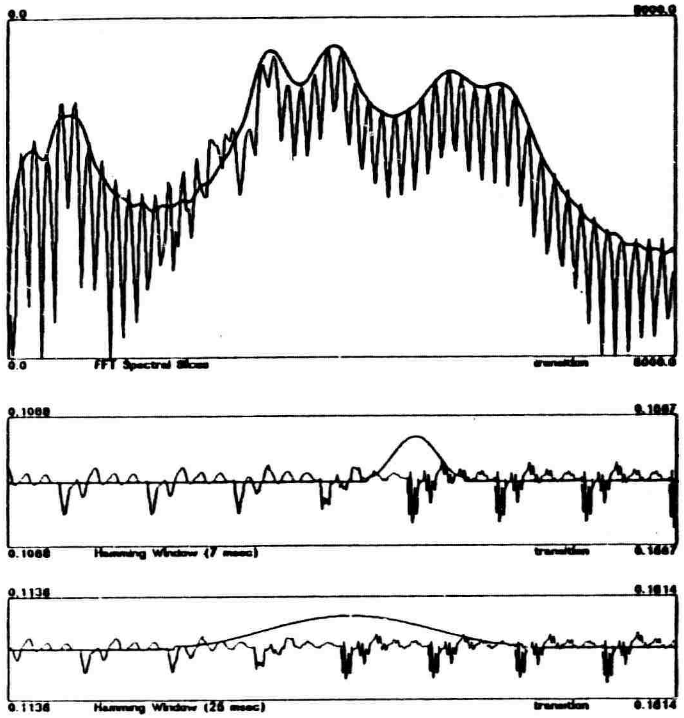


Figure C.6: Hamming Windows Centered at End of Formant Transition

The top display contains DFT spectra for two different duration hamming windows (7 msec and 25 msec). The thick line corresponds to the shorter hamming window. The other displays contain the hamming windows and the original speech waveform.

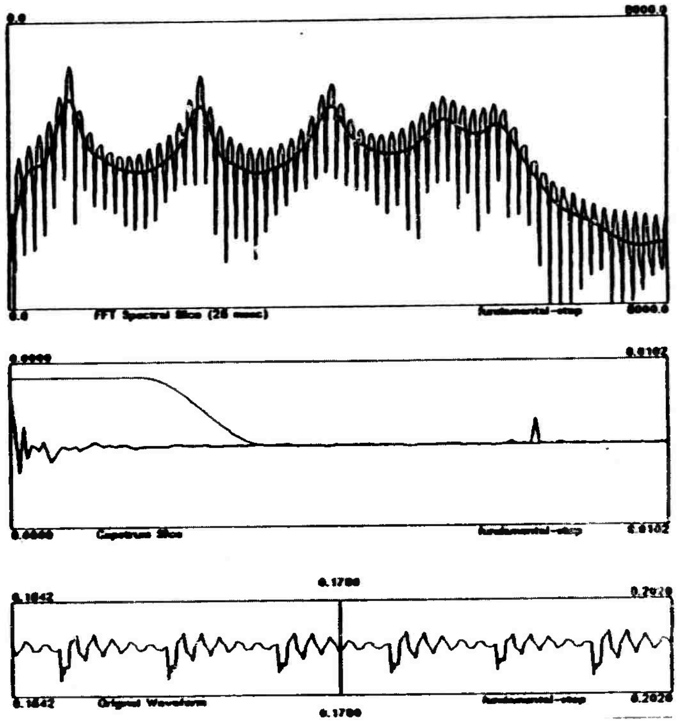


Figure C.7: Cepstrally Smoothed Spectra with Low Pitch Frequency

The top display contains the outputs of unsmoothed and smoothed DFT spectra (hamming window 25 msec). The middle display contains the cepstrum and the smoothing window (2 msec flat, 2 msec raised cosine). The bottom display shows the original speech waveform (pitch approximately 120 Hz).

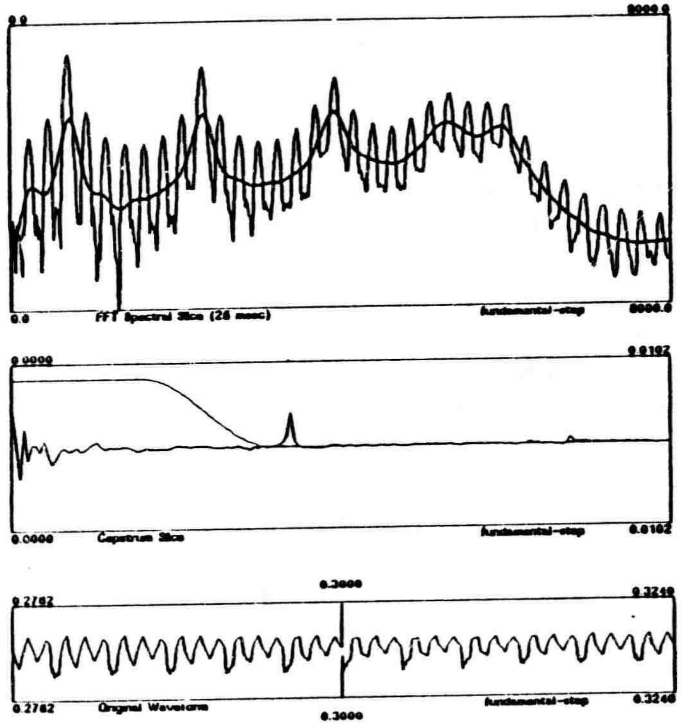


Figure C.8: Cepstrally Smoothed Spectra with Middle Pitch Frequency

The top display contains the outputs of unsmoothed and smoothed DFT spectra (hamming window 25 msec). The middle display contains the cepstrum and the smoothing window (2 msec flat, 2 msec raised cosine). The bottom display shows the original speech waveform (pitch approximately 220 Hz).

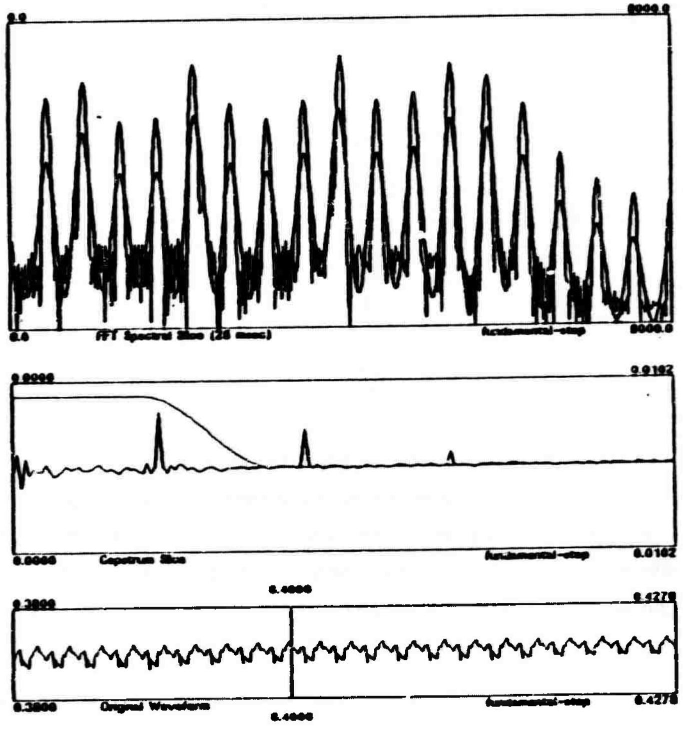


Figure C.9: Cepstrally Smoothed Spectra with High Pitch Frequency

The top display contains the outputs of unsmoothed and smoothed DFT spectra (hamming window 25 msec). The middle display contains the cepstrum and the smoothing window (2 msec flat, 2 msec raised cosine). The bottom display shows the original speech waveform (pitch approximately 400 Hz).

C.2 Vocal Tract Modelling

An alternative to spectral representations based on filter banks or homorphic analysis is to use an approach based on estimating the parameters of a vocal tract model. In fact for most models the vocal tract response $V(z)$ is considered as only one part of the overall frequency response of the speech signal $H(z)$. In general the glottal pulse and radiation components, $G(z)$ and $R(z)$ are taken into acconnt as well.

$$H(z) = G(z)V(z)R(z) \tag{C.3}$$

For this model, the original input is considered to be a train of impulses at the pitch period.

One such model could consist of representing the overall transfer function in terms of a general transfer function of the form

$$H(z) = G\frac{\prod_{i=1}^{q}(z - z_i)}{\prod_{i=1}^{p}(z - z_i)} \tag{C.4}$$

where the parameters used to describe the speech signal are the poles and zeros of the transfer function and the gain factor G. In general, the impulse response associated with the transfer function is a nonlinear function of the numerator and denominator coefficients. Estimating these parameters for a segment of speech would thus typically require the solution of a set of nonlinear equations. For the special case in which the order of the denominator polynomial is zero, the determination of the parameters based on a mean-square error criterion reduces to the solution of a set of linear equations. For the case where the order of the numerator polynomial is zero, the mean-square error criterion reduces to the solution of a set of linear equations of the inverse filter. All-pole modelling is very common for speech analysis and is commonly known as Linear Predictive Coding (LPC) [42], [45].

One important attribute of the vocal tract transfer function is that it is characterized primarily by resonances which are well represented by poles. However, difficulties can arise when the model is invalid, as is true for nasal consonants and nasalized vowels. Figure C.10 shows examples of synthetic stimuli with zeros included at low frequencies. In the top display the zero is located at 1000 Hz as might be found in a nasal consonant. This zero creates a dip in the DFT spectra which is not captured by the LPC spectra. In the bottom display the zero is located at 450 Hz between two poles as might be found in a nasalized vowel. This pole-zero-pole combination is again not captured in the LPC representation although it exists in the DFT spectra. These figures can be compared to cepstrally smoothed spectra as shown in figure C.11 where the essence of the DFT spectra have been captured satisfactorally.

Clearly it is possible to modify the model so that one is better able to match the DFT spectra. For instance in the above examples it is possible to use more poles (19 poles were used for 8000 Hz bandwidth), or to attempt pole-zero modelling. However, no modelling procedure will work correctly all of the time. For this reason it was decided to use a spectral representation such as cepstrally smoothed spectra which does not rely on any underlying model of the speech waveform, and so will tend to be more robust.

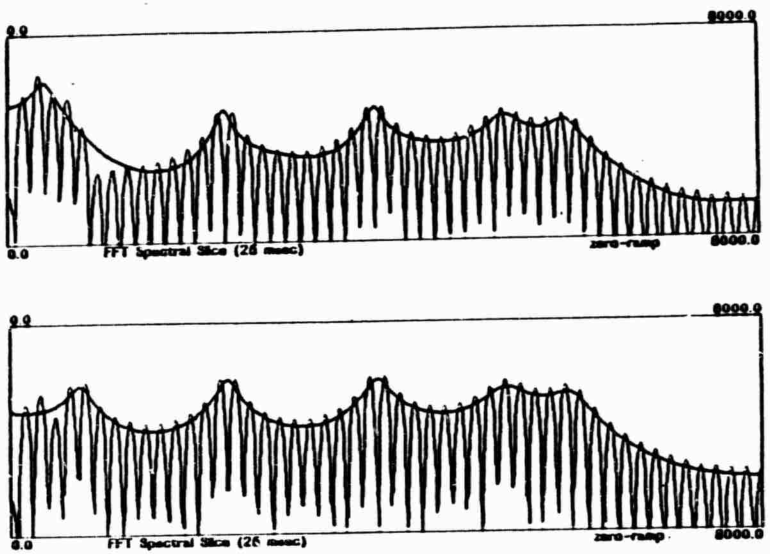


Figure C.10: LPC Spectra

The top display contains the outputs of DFT and LPC spectra (dark line) of a synthetic token containing a zero at 1000 Hz. The bottom display contains the outputs of DFT and LPC spectra (dark line) of a synthetic token containing a zero at 450 Hz. Hamming windows of 25 msec duration were used.

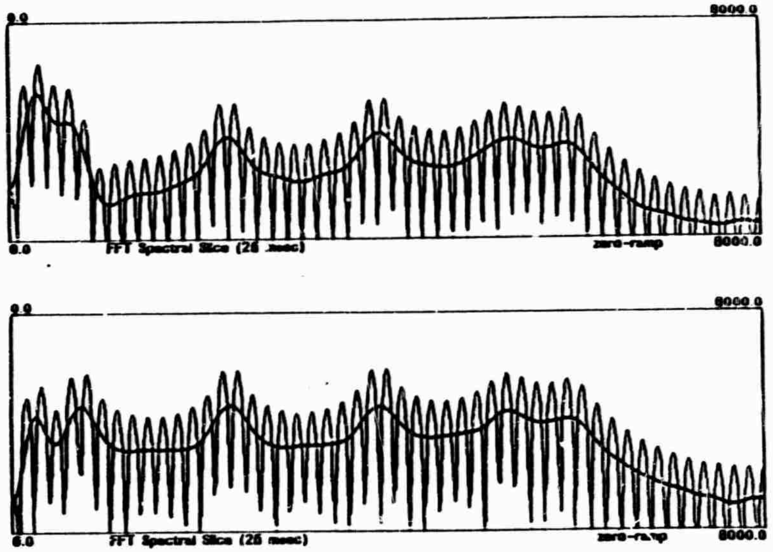


Figure C.11: Cepstrally Smoothed Spectra

The top display contains the outputs of DFT and cepstrally smoothed spectra (dark line) of a synthetic token containing a zero at 1000 Hz. The bottom display contains the outputs of DFT and cepstrally smoothed spectra (dark line) of a synthetic token containing a zero at 450 Hz. Hamming windows of 25 msec duration were used.

Appendix D

Nasalized Vowel Algorithms

Since nasality manifested itself more subtly in nasalized vowels than in nasal consonants, the algorithms used to extract this property from vowels were more sophisticated, and as a result, more fragile, than those used for the nasal consonants. There were two types of calculations used for the analysis of nasalized vowels: those which performed general statistical measures of the spectral distributions, and those which performed peak picking, and measured properties of actual resonances. Both types of computations were found to be effective in distinguishing nasalized vowels from non-nasalized vowels. The following sections describe the algorithms used for each type of calculation.

D.1 Statistical Calculations

Center of Mass

Since the center of mass was found to be a rather ineffective measure of nasalization, its main role was to find the center of energy in the low frequency regions so that the local spread of energy could be measured by the standard deviation calculation.

The center of mass is a specific kind of spectral weighting algorithm, described in chapter 2, where the weighting window, $\vec{W}$, is linear with freqnency. Calculated between two frequency ranges, f_1, and f_2, the center of mass, $\bar{f}$, is defined as

$$\bar{f} = \frac{1}{A_1} \sum_{f=f_1}^{f_2} f X(f) \qquad (D.1)$$

$$A_1 = \sum_{f=f_1}^{f_2} X(f) \qquad (D.2)$$

where $X(f)$, is the value of the DFT spectra at frequency f. For use in nasalized vowels, the center of mass was computed between 0 and 1000 Hz, which covers the first formant range of most men and women [59].

In order to reduce the sensitivity of the center of mass function to sudden changes at the end points, such as a formant passing below 1000 Hz, the DFT spectra was windowed with a trapezoidal window before the center of mass was computed. The window was flat between 100 Hz and 900 Hz, and had 100 Hz tapers at each end. Windowing the spectra ensured that there were no sudden changes in the center of mass caused by a marginal movement in energy across the upper boundary.

There are several different spectral representations on which the center of mass could have been computed (magnitude squared, or magnitude for instance). However, the log magnitude squared (dB) spectrum was used because it was observed that the extra resonance frequency had the largest effect on the center of mass in this representation. In any other representation, the major resonance peak dominated the value of the center of mass.

Using the log spectrum introduced a sensitivity problem into the calculation however. In a magnitude spectra the baseline value for the center of mass is zero. There is no such corresponding baseline valne for the dB scale however since values may go to $-\infty$. Thus, some form of normalization is necessary. Typical normalization procedures establish some baseline value, relative to a value in the spectrum. Note that the center of mass may be made arbitrarily sensitive this way. In this research, a good value was found to be somewhere around 20 dB below the spectral peak in the frequency range of interest. This yielded a center of mass which was responsive to changes in the first formant frequency and nasality in the vowel, but was not overly sensitive to minute changes in the spectrum.

Standard Deviation

A measure of the local spread of energy around the center of mass was found to be a very good measure of nasalization. This was calculated by measuring the second moment of local energy around the center of mass. The term "local" was defined to include all energy within a specified frequency radius of the center of mass. Thus, if the center of mass was measured to be 700 Hz, and the frequency radius, f_r, had been defined as 200 Hz, then the standard deviation would be calculated between 500 and 900 Hz. In general, the standard deviation, σ, is calculated between $\bar{f} - f_r$, and $\bar{f} + f_r$, and is defined as

$$\sigma = \sqrt{\frac{1}{A_2} \sum_{f=\bar{f}-f_r}^{\bar{f}+f_r} X(f)(f-\bar{f})^2} \qquad (D.3)$$

$$A_2 = \sum_{f=\bar{f}-f_r}^{\bar{f}+f_r} X(f) \qquad (D.4)$$

The same issues which were discussed for center of mass apply here. Thus the standard deviation was computed on the same normalized log magnitude spectrum which was used to calculate the center of mass.

The frequency range is a very important parameter since it determines the type of deviation that is being measured. The most effective range was found to be 500 Hz on either side of the center of mass. Thus, the standard deviation was measuring the overall spread of energy in the low frequency region, rather than the local spread of the first formant.

Since the center of mass was measured between the ranges of 0 to 1000 Hz, at least one end point in the standard deviation calculation would extend outside the center of mass endpoints (unless the center of mass was exactly 500 Hz). In order to include energy outside the first formant region, which was detrimental to the standard deviation measure, the standard deviation only in the valid regions. In other words, if the center of mass was 700 Hz, the standard deviation was computed between 200 Hz and 1000 Hz.

Although the frequency range restriction was necessary, it made the value of the standard deviation measure frequency dependent. In fact, the maximum valus of the standard deviation at any frequency, would be linearly related to the width of the frequency region used in the calculation. Thus, if a deviation value was computed over an 800 Hz range, its maximum value could only he 0.8 that of a deviation which used a full 1000 Hz range. In an attempt to normalize the standard deviation, so that it was frequency independent, each value was scaled upwards by the ratio of the maximum frequency width (1000) to the actual frequency width used in the calculation. This procedure was found to substantially reduce the frequency dependence of the standard deviation calculation.

D.2 Resonance Calculations

Qualitative observations indicated that it would be useful to measure certain properties of the actual resonances in the low frequency region of the spectra. Before this could be done , the resonances themselves had to be found. To do this, spectral regions were established by searching for zero crossings in the second derivative of the smoothed log spectra, as was illustrated in chapter 2. Once the spectral regions were established, resonances could be found by collecting all the peak regions below 1100 Hz.

The actual collection algorithm only gathered resonances until it either had two, since one would be a first resonance and the other a nasal resonance, or until it had passed over 1000 Hz. Note that this procedure introduces a flaw into the system, since non-nasalized high back vowels could be collected if the second formant was below 1000 Hz. The magnitude of this problem was reduced by checking to make sure that if there were two peaks collected, one of them was actually below 400 Hz. This ensured that at least one resonance was either a nasal resonance, or a very low first formant. Thus, if two resonances were found at 500 Hz and 800 Hz, the 800 Hz resonance would be rejected, and the 500 Hz resonance would be kept. Thus, the main vowel which caused problems with this sorting algorithm was /u/.

Percentage

Once the two lowest resonances were established, the percentage measure was calculated in a given time region by dividing the number of spectral slices which had two resonances in the time region, by the number of spectral slices in the time region.

Resonance Dip

Whenever there were two resonances in the spectrum, the resonance dip was calculated by measuring the difference, in dB, between the smallest resonance, and the valley. If there were not two resonances. no value of resonance dip was computed.

Resonance Difference

Whenever there were two resonances in the spectrum, the resonance difference was calculated by measuring the difference, in dB, between the second resonance, and the first resonance. Note that no attempt was made to determine which resonance was the nasal resonance. Thus for high vowels, the resonance difference was negative, and for low vowels, the resonance difference was positive. If there were not two resonances, no value of resonance difference was computed.

A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech

by

Hong Chung Leung
B.E.E.E., City College of New York
(1981)

Submitted in Partial Fulfillment
of the Requirements for the
Degrees of

Master of Science

and

Electrical Engineer

at the

Massachusetts Institute of Technology

January 1985

Signature of Author ....
Department of Electrical Engineering and Computer Science
January 10, 1985

Certified by ....
Victor W. Zue
Thesis Supervisor

Accepted by ....
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech

by

Hong Chung Leung

Submitted to the Department of Electrical Engineering and Computer Science on January 10, 1985 in partial fulfillment of the requirements for the degrees of Master of Science and Electrical Engineer.

Abstract

This thesis is concerned with the design and implementation of a system for automatic alignment of phonetic transcriptions with continuous speech. The motivation for this work stems from the fact that a large database of time-aligned speech material is necessary for speech researchers to quantify the acoustic properties of different speech sounds, and the fact that manual alignment is tedious and unreliable. The implemented system consists of three modules. The speech signal is first segmented into broad classes using a non-parametric pattern classifier. Path finding techniques are then used to align the broad classes with the phonetic transcriptions. These aligned broad classes provide "islands of reliability" for more detailed segmentation and refinement of boundaries. Specific speech knowledge is utilized throughout the system. By doing alignment at the phonetic level, the system can often tolerate inter and intra-speaker variability. The system was evaluated on one hundred sentences, spoken by three male and two female speakers. 97% of the segments are mapped into only one phonetic event, 75% of the time the offset between the boundary found by the automatic alignment system and a trained transcriber is less than 10 ms. Supporting software has also been developed so that final manual adjustments, if needed, can be made.

Thesis Advisor: Victor W. Zue
Title: Assistant Professor of Electrical Engineering and Computer Science

In memory of my mother, and my brother, Chi.

Acknowledgments

There are several people that I would like to thank for their contributions to this research. My biggest debt of gratitude goes to my thesis advisor Victor Zue for providing me with a model for how to conduct research, guiding me in the right directions, and for providing an outstanding research environment.

I would like to thank the members of the Speech Communication Group for their input. In particular, I thank D. Scott Cyphers for improving SPIRE, maintaining the lisp machines, and for his friendship. I also thank Francine Chen and Lori Lamel for teaching me how to read spectrograms.

My deepest appreciation is for my family. I would like to thank my parents for instilling within me the quest for excellence, and my sister for her love and care.

Finally, I would like to thank Suk for her endless love, encouragement and support, especially during my most difficult times in this country.

Table of Contents

Chapter One

Introduction

1.1 Problem Definition

This thesis deals with the development of an automatic procedure for alignment of the speech signal with its corresponding phonetic transcription. Figure 1-1 illustrates an acoustic signal and the corresponding phonetic transcription of the phrase, "Glue the sheet to the dark..". The transcription is determined by an acoustic phonetician, who listens to the speech signal and visually examines its various displays. The arrows in the Figure are drawn manually at the acoustic landmarks that correspond to the phonetic transitions. The objective of this thesis is to automate such a process. In other words, given the speech signal and the phonetic transcription, the goal is to automatically locate the acoustic boundaries of these phonetic events as robustly and consistently as possible.

At the phonetic level, a speech utterance consists of a sequence of phonemes. However, due to the interaction among the various articulatory structures and their different degrees of sluggishness, the acoustic cues of adjacent phonemes overlap with each other. Thus locating the absolute phonetic boundaries in a continuous acoustic speech signal is a very difficult, if not impossible, goal to achieve. A more realistic goal is to locate the acoustic landmarks that correspond to the phonetic transcription consistently and robustly. This is the objective of the thesis.

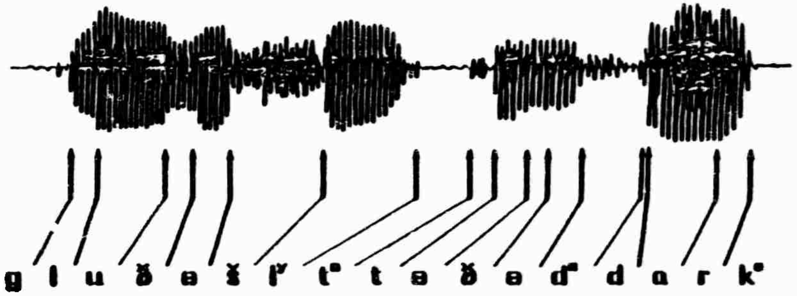


Figure 1-1: Alignment of the speech signal with the phonetic transcription

The reliability of the acoustic landmarks in continuous speech is not at all uniform. There is a continuum of difficulty. Some landmarks are obvious and clear while others are more subtle. Figure 1-2 illustrates the spectrogram and various displays for the same phrase, "Glue the sheet to the dark.." The phonetic transcription, which is shown above the spectrogram, is manually aligned by an experienced acoustic phonetician. As can be seen in the Figure, the transition from a strong fricative to a vowel, as in the word "sheet", is strongly evidenced by the abrupt decrease of high frequency energy and a sharp onset of low frequency energy. This kind of acoustic landmark is relatively easy to detect. On the other hand, the transition between various similar sounds may be more subtle. For example, the transition from a vowel to a post-vocalic /r/ as in the word "dark" is not marked by any distinct acoustic cues. This kind of acoustic landmark is, in general, difficult to locate without first establishing the phonetic context.

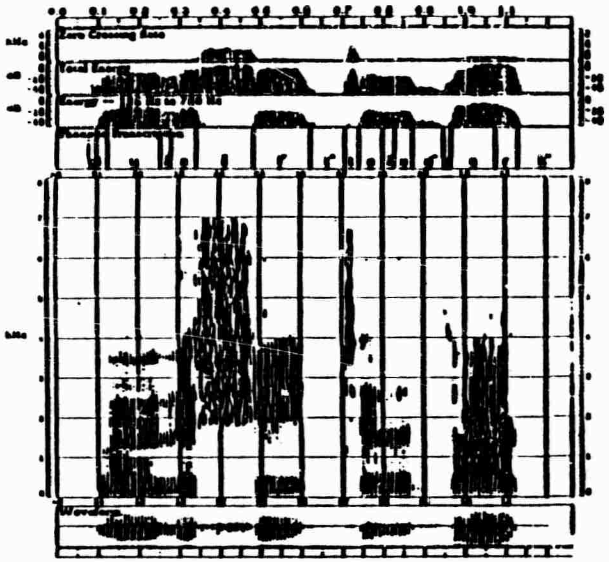


Figure 1-2: Spectrogram and various displays for the phrase, "Glue the sheet to the dark.."

1.2 Motivation

It has long been recognized that the relationship between phonetic segments and their acoustic realizations is very complex. Therefore, studies of this relationship must be conducted with a sufficiently large database of speech material. However,

in order to enable the user to have direct access to specific portions of the speech signal, the utterances must be augmented with a set of time-aligned transcriptions. The time-aligned transcriptions can serve as pointers to specific phonetic events in the speech signal. If a sufficient amount of time-aligned acoustic data is available, speech researchers will then be able to quantify the properties of phonetic segments and describe how their characteristics are modified by contexts. Thus, for example, researchers will be able to query the database for all occurances of pre-vocalic and post-vocalic /t/ and quantify the similarities and differences between them. They can also study different speech phenomena, such as structural constraints, coarticulation, phonological processes, and speaker characteristics.

A large database of aligned speech material is particularly important for phonetic recognition. In addition to helping researchers to develop better acoustic-phonetic rules for higher recognition performance, automatic time-alignment can also serve as a testbed for phonetic recognition. It is well known that classifying the speech signal into detailed segments is a very difficult task [Zue 80]. By knowing what the phonetic events are, it should be relatively easier to find the detailed phonetic segments. Thus, automatic time-alignment can also be viewed as a process of locating specific phonetic segments when the identity and the context of the segments are known. It is a learning step towards phonetic recognition and it can also enable evaluation of a specific algorithm.

Traditionally, the alignment is performed manually by a trained acoustic-phonetician, who listens to the speech signal and visually examines various signal displays. There are several disadvantages to this approach. First, the task is extremely time consuming; even under the best of circumstances, the process of time alignment can take several minutes for one second of speech material. Second, the task requires the skill and knowledge possessed by a small number of experts. These two reasons combine to severely limit the amount of transcription aligned data that can be collected in this manner. Third, there is the lack of consistency and reproducibility of the results. Manual labeling often involves decisions that are highly subjective. Even if the sentence and the transcription were the same, the inter- and intra-transcriber variability can still be quite high. Finally, there is the problem of human error associated with tedious work.

The problem associated with manual labeling, together with the need for a large corpus of time-aligned speech data, clearly call for the development of an automatic time-alignment system. With such an automatic system, it would be much easier for speech researchers to establish a large acoustic-phonetic database. The time-alignment system should also be speaker-independent so that a database with multiple speakers can easily be established.

1.3 Review of Literature

In doing automatic alignment of phonetic transcriptions with continuous speech, both the digitized speech signal and the phonetic transcriptions are given. They are

two different levels of representations of an utterance. The phonetic transcription is a symbolic representation of an utterance and it contains linguistic information about the utterance. The speech signal, however, is an analog signal and it contains linguistic and extra-linguistic information. In order to time-align the phonetic transcription with the speech signal, these two representations must be brought to the same level and the extra-linguistic information must also be properly dealt with.

Over the past few years, several automatic time alignment procedures have been suggested in the literature. There are, in general, two different approaches. One approach is to generate a reference utterance from the transcription and to perform the time-alignment on a frame-by-frame basis, using dynamic programming algorithms. There are, in general, three methods to obtain the reference utterance. One is to previously label an utterance of the same orthographic transcription. A dynamic programming algorithm can then match the labels to the other utterance. Chamberlain and Bridle suggests ZIP, a modified dynamic programming algorithm designed to compute the time alignment of two utterances of the same text [Chamberlain 83]. Hohne et al. also suggest a similar approach by utilizing an unconstrained endpoint, local minimum dynamic time warping algorithm [Hohne 83]. The second method is to synthetically generate an utterance of the same phonetic transcription [Lennig 83]. The third method is to match the speech signal with a concatenation of stored templates [Lowry 78]. In order for this first approach of generating a reference utterance to be effective, the two utterances must not differ significantly in detailed phonetic structures, and the synthesis rules must be fairly advanced. The alignment procedure must also be able to deal with the variability due to different speakers and coarticulatory effects.

Another approach is to divide or classify the speech signal into segments which correspond to relevant phonetic events. A dynamic programming algorithm can then be used to align these two sequences of symbols, namely, the phonetic transcription and the labeled segments of the utterance. Wagner suggests a system which allows the direct mapping of a phonetic transcription onto an acoustic parameter representation of continuous speech [Wagner 81]. Linear prediction analysis, segmentation and formant tracking provide the acoustic parameters for voiced, unvoiced and silent segments. The given phonetic transcription is expanded to include implicit phone sequences and transitions. Labeling is then performed in two stages using dynamic programming. Segment labeling is achieved by mapping the expanded phone string onto the acoustic segments using a dynamic programming algorithm. The distance functions are determined heuristically. A more detailed frame-by-frame labeling is then achieved by a second dynamic programming algorithm, using derivatives of energy and formant functions. This approach has the advantage that no reference utterance is needed. However, the success of this technique is closely tied to the proper utilization of acoustic-phonetic knowledge.

There is one major difference between these two approaches. In the first approach, the acoustic landmarks are obtained as by-products of the optimal path. While this path may be optimal in the global sense, the landmarks so obtained may not be located at places where strong acoustic evidence can be observed. This is because the path is heavily weighted by the long and steady portions of the speech signal. Thus the second approach has the advantage that more attention can be placed at the acoustic landmarks.

1.4 Overview of the Thesis

Current state of the art in automatic alignment of the speech signal with its corresponding phonetic transcription relies heavily on dynamic time-warping techniques. While the warping path may be optimal in some global sense, the acoustic landmarks so obtained may not correspond to places where strong acoustic evidence can be observed. Furthermore, such an approach has the difficulty dealing with the highly variable acoustic signal. In this thesis, a new approach to automatic alignment is proposed, designed, and implemented. Rather than time-warping two utterances on a frame-by-frame basis, the speech signal is first pre-processed by an initial broad classifier. The broad class segments so produced can then provide "islands of reliability" for more detailed alignment.

In Chapter 2, some of the acoustic and phonetic properties of speech sounds are discussed. The variability of the speech signal results from local phonetic context and from different speaker characteristics. These multiple sources of variability makes it difficult to locate the detailed phonetic events in a context-independent manner. This suggests pre-processing the speech signal by a broad class segmenter. The output of the segmenter can then establish the phonetic context. As a result, specific context-dependent procedures can be applied. Proper utilization of speech knowledge will also be discussed. A feasibility study of using broad class descriptions will also be described.

Chapter 3 presents description of the alignment system. An outline of the system is presented and each of the components is subsequently discussed in detail. The initial broad classification is performed by using traditional pattern classification techniques, augmented with acoustic-phonetic knowledge. The broad segments so produced are then aligned with the phonetic transcription by utilizing path finding techniques. Finally, the more detailed phonetic events are located by applying specific algorithms and features.

In Chapter 4, we show the results of the system evaluation and performance. The evaluation is broken down into the separate components of the system. Comparison with manual alignment is also given.

Chapter 5 presents a summary of the thesis. In addition, suggestions for further research are discussed.

Chapter Two

Design Considerations

2.1 Variability in the Speech Signal

Although the phonemes may be characterized by a set of invariant and distinct acoustic features, the acoustic speech signal has a very high degree of variability. There are three major factors. First, the acoustic realization and characteristics of a phoneme can be modified severely as a consequence of immediate phonetic context. When spoken continuously and naturally, a sequence of phonemes bear little acoustic resemblance to the phonemes uttered in isolation. As can be seen in Figure 2-1 the phoneme /t/, for example, can have many different acoustic patterns. Second, the speech signal contains a good deal of extra-linguistic information, such as background noise, the emotional and physiological states of the speaker, the type of speaking environment, etc. In doing alignment with the phonetic transcription, such extra-linguistic information is not necessary and is, in fact, undesirable. Finally, different speaker has different realizations of a phoneme. Figure 2-2 illustrates spectrograms for the same phrase, "Glue the sheet to the dark..", spoken by two different speakers, one male and one female. As can be seen in the Figure, the two utterances have very different acoustic realizations, especially the intervocalic /ð/'s, a voiced weak fricative. Thus, it is a difficult task to locate the detailed phonetic events in a continuous speech signal.

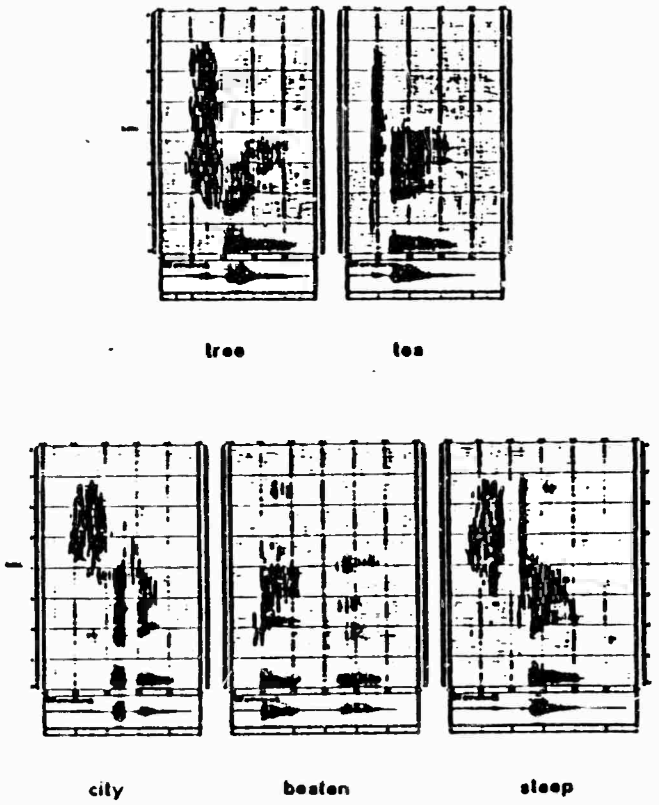


Figure 2-1: Spectrograms illustrating the acoustic realizations of the various allophones of /t/.

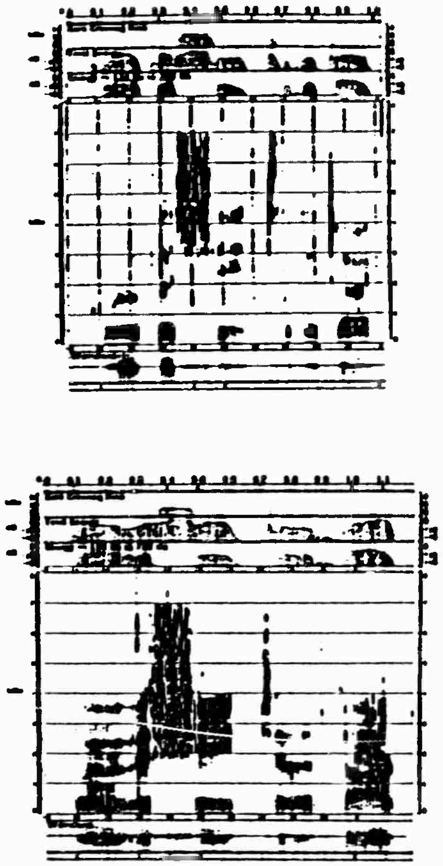


Figure 2-2: Spectrograms illustrating different speaker characteristics for the phrase, "Glue the sheet to the dark.."

Although the speech signal is highly variable due to a number of factors, much of the variability are actually regular with the phonetic context. For example, although the acoustic realizations of the phoneme, /ð/, in Figure 2-2 are quite different, it is expected that the phoneme in an intervocalic context can have realization like that in the second spectrogram. Therefore, if the context has been established, locating a specific phonetic event can be much easier. Furthermore, some of the acoustic landmarks are relatively robust and can be captured reliably without using context. Consequently, one can first locate these robust acoustic landmarks, and use them to establish the context for more detailed alignment. In order to perform the detailed alignment efficiently, the context should be as specific as possible. However, the context must also be broad for it to be reliable.

2.2 Application of Speech Knowledge

By using such a context-independent approach, specific speech knowledge can be applied directly at different stages. It can be utilized to locate the robust acoustic landmarks, which can then be used to establish the context. By knowing what the phonetic context is, speech knowledge can then be utilized to locate the specific phonetic events. Thus, for example, the strong fricative, /š/, in Figure 1-2 can be located by detecting the presence of a large amount of high frequency energy. After the strong fricative has been located, the time span of the first five phonetic events can then be found. Since the voiced weak fricative is inter-vocalic, it can be located

by applying specific algorithms and feature parameters. Other phonetic events can also be located in a similar manner. Therefore, by selecting the proper set of algorithms and extracting the proper set of feature parameters, one may expect that the alignment can perform better than if the speech signal is represented by some uniform and arbitrary measure, such as the short-time spectrum or the linear predictive coefficients. In addition, by using appropriate feature parameters that characterize the phoneme, the influence of the extra-linguistic "noise" can be minimized. Thus, this approach has the potential for performing better than other context-independent methods.

The importance of applying specific speech knowledge is well demonstrated by FEATURE, a speech recognition system developed by Cole et al [Cole 83]. In this approach, the classification of the speech signal is performed by applying Fisher linear discriminant analysis and multivariate Bayesian analysis assuming Gaussian probability densities of the features. About fifty acoustic features are utilized by the system, and they are discovered through examination of various visual displays of the speech signal. Measured in terms of performance, the FEATURE system offers the best recognition results in the literature for the task of alphabetic recognition. Furthermore, it proves that a system with proper utilization of speech knowledge can out-perform systems that does not apply the knowledge.

2.3 Islands of Reliability: Broad Class Description

In order to establish the phonetic context, the speech signal must be somehow segmented into sections. The varying degrees of difficulty in locating the acoustic landmarks suggests that broad phonetic characterizations may be used to establish the robust phonetic context. Furthermore, studies examining the confusability of English phonemes indicates a strong grouping of confusions within broad classes based on manner of articulation differences [Miller 54]. Manner of articulation refers to the method of production of a particular phoneme. For example, the phonemes /s/ and /š/ are produced in the same manner, with a constriction somewhere in the vocal tract and the excitation acting as a noise source. These fricatives, which share manner of articulation, differ mainly in their respective place of articulation, or the position of the constrictions. Due to the speech production process, manner of articulation differences tend to have more robust and speaker-invariant acoustic cues than place differences. This makes using broad manner class description as islands of reliability attractive for performing alignment.

This approach of using broad manner classes as islands of reliability has a further advantage. As previously mentioned, most of the automatic alignment procedures suggested in the literature rely on some kind of frame-by-frame time-warping with a pre-determined reference utterance. While there are difficulties in dealing with the acoustic and phonetic variability in the speech signal, the globally optimal path does not necessarily locate the phonetic events reliably. This optimal path is heavily

weighted by the long and steady state portions of the utterance while the transitional regions, which are usually short in duration, are very lightly weighted. This is an undesirable characteristic because we are actually most interested in the transitions between the phonetic events. This suggests that the speech signal should first be passed through a pre-processor or, more specifically, a segmenter, which focuses on locating the phonetic transitions or the acoustic landmarks reliably. The output of the pre-processor can then be used to perform the actual alignment.

In order to use the broad class segments as "islands of reliability", the broad representation must be aligned with the phonetic transcription. Since the broad segments carry important acoustic information about the speech signal and the transcriptions carry important linguistic information about the same signal, the alignment can be performed with appropriate application of acoustic-phonetic knowledge. This is, again, in contrast with the frame-by-frame time-warping approach.

2.3.1 Previous Applications of Broad Classes to Speech Recognition

The approach of using robust acoustic-phonetic events as anchor points for automatic alignment is in agreement with the broad class representation suggested in the speech recognition literature. In 1982, Shipman and Zue presented results demonstrating the powerful constraints imposed by the sound patterns of American English [Shipman 82]. They showed that the broad phonetic representation of any word matches, on the average, only a small number of the words in a 20,000 word lexicon. Thus, they proposed that speech recognition be performed by initially classifying the speech signal in terms of a broad phonetic sequence. This broad phonetic sequence can then be used for lexical access to limit the number of word candidates. Finally, more detailed acoustic-phonetic analysis can be performed to determine which of the word candidates was actually spoken.

This isolated word recognition approach was then further studied by Huttenlocher [Huttenlocher 84]. He investigated how variability in the speech signal affects the broad phonetic representation of a word and demonstrated that the redundancy in the lexicon is sufficient to allow for reasonable recognition errors, without substantially degrading the performance.

The same approach was also pursued by Chen in continuous digit recognition [Chen 84]. Although there are usually many more phonetic events in a continuous sentence, the number of word candidates can still be limited to a small number, due to the small size of the vocabulary. Chen demonstrated that by using an ideal broad class representation, 70% of the correct word boundaries can be uniquely identified. Although the performance may degrade with the use of a realistic broad classifier, her study does show that the allophonic constraints in continuous speech can be exploited with a limited vocabulary size.

2.4 A Feasibility Study

All the above results demonstrate that the broad category classes can provide powerful constraints for speech recognition. The question now is: Can these broad classes provide enough constraints for alignment?

A feasibility study was conducted to answer this question. One hundred sentences spoken by three male and two female were used in the study. The sentences were selected from the Harvard list of phonetically balanced sentences. There is approximately 4 minutes of speech material. Figure 2-3 summaries the total number of the different broad phonetic classes in the sentences. There are altogether about 2700 phonetic events.

All sentences were manually time-aligned with the corresponding phonetic transcriptions by an experienced acoustic-phonetician. An "ideal" segmentation with 5 broad classes, derived from the time-aligned phonetic transcriptions, is then associated with each sentence. To make the "ideal" segmentation correspond more closely to the output of a realistic segmentation, segments of different phonetic classes may be assigned the same label. For example, although liquids and vowels are two different phonetic classes, their acoustic realizations are quite similar. This similarity makes it quite difficult for a broad classifier to distinguish the two classes. Therefore, they are both assigned a 'sonorant' label in the ideal segmentation. Furthermore, adjacent segments with the same label are collapsed into one long segment with the same label. Thus, although the segmentation is ideal, it retains the

BROAD PHONETIC CLASS	COUNT
VOWEL	868
NASAL	196
LIQUID/GLIDE	276
STOP	828
AFFRICATE	31
STRONG FRICATIVE	240
WEAK FRICATIVE	220
ASPIRATION	25
TOTAL	2684

Figure 2-3: Distribution of the broad phonetic classes in the database.

general characteristics of a realistic broad class segmentation algorithm. Figure 2-4 summarizes the mappings between the phonetic classes and the broad acoustic labels. The labels are: S (vowel-like sonorant), O (obstruent), B (nasal and voice-bar), - (silence) and D (energy dip in sonorant region). Since the mappings are context-independent, the segmentation so obtained is rather ideal. For example, the

phoneme /v/ is always mapped into obstruent, although an intervocalic /v/ may have an acoustic realization very different from an obstruent. Nevertheless, this ideal segmentation provides an upper bound to the performance of a realistic broad class segmentation.

PHONETIC CLASS	IDEAL BROAD CLASS LABEL
VOWEL, LIQUID, GLIDE	S
FRICATIVE, STOP RELEASE	O
NASAL VOICED STOP CLOSURE	B
SILENCE UNVOICED STOP CLOSURE	-
FLAP	D

Figure 2-4: Mappings of the broad phonetic classes into the segmentation labels.

Figure 2-5 summarizes statistics on the number of phonetic events in one broad class segment. 85% of the time there is only one phonetic event in a segment, whereas 12% of the time there are two phonetic events in a segment. In other words, the broad class segmentation can divide an utterance into a sequence of units, 97% of which correspond to only one or two phonetic events.

SEGMENT LABEL	1	2	3	4	5	TOTAL
S	616 (67%)	183 (24%)	42 (5%)	28 (4%)	4 -	773
O	736 (91%)	65 (8%)	6 (1%)	0 -	0 -	807
B	289 (90%)	34 (10%)	1 -	0 -	0 -	324
-	484 (100%)	0 -	0 -	0 -	0 -	484
D	63 (100%)	0 -	0 -	0 -	0 -	63
TOTAL	2078 (85%)	282 (12%)	49 (2%)	28 (1%)	4 -	2441

Figure 2-5: Statistics on number of phonetic events in one segment produced by the ideal segmentation

These two sequences of symbols, namely, broad segment labels and phonetic transcription, are then time-aligned by using a path finding algorithm to be described later. To simulate a real automatic time-alignment system, the time marks of the phonetic transcription is not used. Thus it is a path finding problem with two sequences of symbols, one time-registered with the waveform and the other is not. After the alignment is performed, the time marks of the phonetic transcriptions, which are obtained from the ideal segmentation, are then used to check if the

phonetic events are mapped into the right broad segment. The result of the experiment is encouraging. Over 99% of the phonetic events are mapped to the right segment. This experiment indicates that a sequence of broad class segments can provide enough constraints to time-align an utterance with its phonetic transcription. It also indicates that with such segmentation, approximately 15% of the total number of segments need to be further processed, since 85% of them correspond to only one phonetic event.

2.5 Chapter Summary

- Finding reliable cues for locating detailed phonetic events is difficult, particularly across a number of speakers and a variety of phonetic contexts.

- Much of the variability are actually regular with context.

- Alignment with first establishing broad phonetic context can be much more effective and reliable.

- With proper utilization of speech knowledge, system performance can be much better.

- Broad phonetic classes based on manner of articulation differences are useful descriptions of the speech signal and can also provide "islands of reliability".

- Broad manner classes are also relatively invariant across speakers and phonetic contexts.

- A feasibility study illustrates that a sequence of broad manner classes can provide enough constraints to time-align an utterance with its corresponding phonetic transcription.

- With the phonetic context established, alignment of detailed phonetic events can be more specific.

Chapter Three

System Description

3.1 System Structure

Figure 3-1 shows the basic structure of the system. The system consists of three modules. First, the speech signal is pre-processed by a broad manner class segmenter. This segmenter detects and locates the major acoustic landmarks in the speech signal. Second, the broad class segments produced at the first stage are aligned with the phonetic transcriptions using path finding techniques. Finally, depending upon the phonetic context, specific algorithms and features can be used to locate the more subtle acoustic landmarks which have not been found at the first stage.

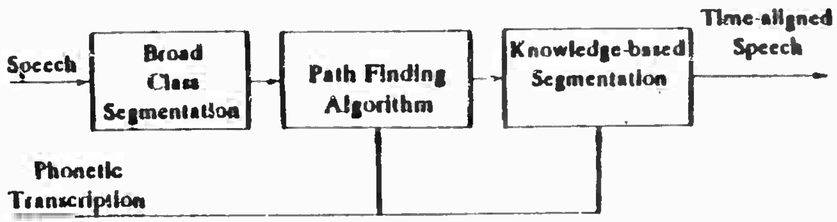


Figure 3-1: Basic system structure

There are several advantages of using such a system structure. The initial classifier can first determine the robust acoustic events that are relatively context independent. Since the segmentation is performed independently of the phonetic transcriptions, it can actually be directly applied to an automatic phonetic recognizer. At the second stage, since the search space is very rich in acoustic-phonetic information, the path finding algorithm can be improved by applying specific acoustic-phonetic knowledge. The output of the alignment can then establish "islands of reliability" for detecting phonetic events that are more context-dependent. Context-dependent algorithms and feature parameters can be used.

3.2 Initial Broad Classification

In order to determine the robust acoustic events in the speech signal, the initial classifier must be able to solve two basic problems. First, it must be able to detect and locate obvious acoustic landmarks. This problem requires using appropriate sets of feature parameters, and deciding whether a landmark exists based on these parameters. Since feature parameters change at different rates across a phonetic transition over a fairly long interval, the classifier must be able to resolve the non-simultaneity. Second, the classifier must label or specify a mapping from acoustic patterns to discrete units, wherein a large number of acoustic patterns map into the same unit. These two problems are made more difficult due to acoustic differences between speakers, and to coarticulation. In some cases, the inter-speaker acoustic

differences can overshadow the intra-speaker differences. For instance the vowels overlap greatly with one another across speakers [Peterson 52]. On the other hand, a given phoneme produced by a given speaker is still significantly affected by coarticulation such that many surface realization may occur.

3.2.1 Decision Structure

There are at least two ways in which the task of making a M-category classification can be structured. The M-category decision can be obtained by a single classifier which assigns one of the M labels to a speech segment, or by a sequence of binary decisions. There are a number of advantages of using a sequence of binary decisions. One potential advantage is that the feature sets that are used for each discrimination can be selected independently. Thus different feature vectors can be used for each classifier to maximize the contrasts between the two possible output classes. For example, zero-crossing rate is used for distinguishing sonorants and obstruents, but not for distinguishing vowels from other voiced consonants.

A second advantage of using a sequence of binary classifiers is that it can allow a more flexible division of the feature space. In a single M-category classifier, each class i is represented by a discriminant function d(i). Thus if x denotes a point in the feature space, then the class i region in the feature space is the set of points for which $d(i,x) < d(j,x)$, for all j not equal to i. As can be seen in Figure 3-2, if a single classifier is used, a highly nonlinear function is necessary to discriminate the two classes in the 2-dimensional feature space. However, if a sequence of binary classifiers is used and each classifier performs a linear discrimination, then the nonlinear characteristic can be approximated by a set of piecewise linear functions as shown in Figure 3-3.

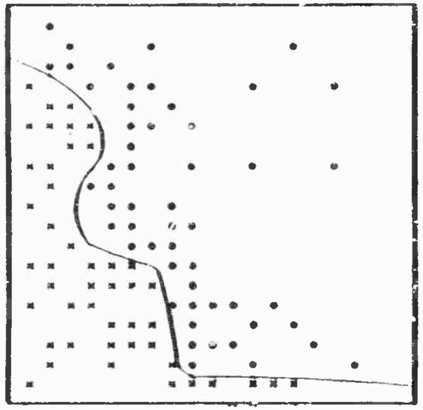


Figure 3-2: A nonlinear surface is necessary to partition the two classes

Thus, a sequence of binary decisions is used as the decision structure in this thesis. However, there is one potential disadvantage. A pattern assigned incorrectly to a wrong class early in the decision tree may never be classified correctly. In order to avoid this kind of error propagation, two classes of data from different nodes are

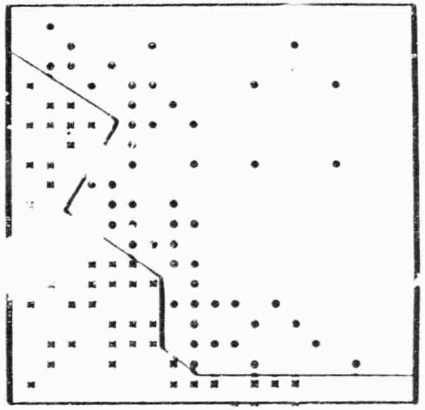
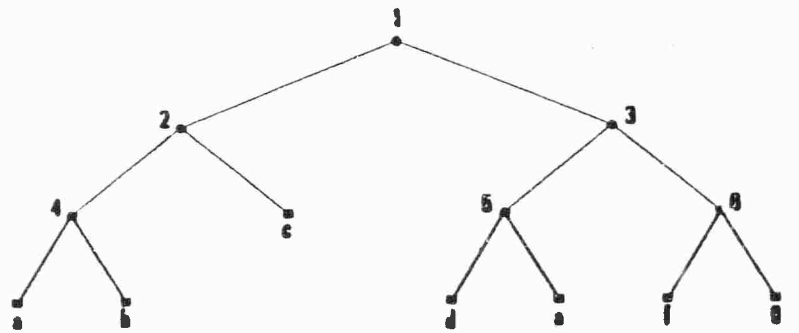


Figure 3-3: A piecewise linear approximation partitions the two classes

allowed to recombine. Therefore, the problem of classifying the speech signal into different robust classes can be reduced to a sequence of sub-problems, each of which is relatively easier to tackle. This approach can also be viewed as a problem reduction representation.

The structure of the binary decision tree is shown in Figure 3-4. The filled circles denote non-terminal nodes and the filled squares denote the terminal nodes. At the top of the decision tree, a binary classifier assigns one of two labels to every frame of speech data. Each group of data will then pass through a different classifier at a lower node, and the process repeats. Note that more than one terminal node can have the same label. The labels of the terminal nodes are the same as those used in the feasibility study. The feature vectors and their dimensions can be different at the decision nodes. The dimensions of the feature vectors are between two and five.



				Label
1.	Speech input	a.	Vowel-like sonorant	S
2.	Sonorant-like	b.	Voiced Consonant	D
3.	Non-sonorant-like	c.	Nasal and voice bar	B
4.	Sonorant-like	d.	Fricative	O
5.	Fricative-like	e.	Silence-like	-
6.	Weak fricative, silence	f.	Weak fricative	O
		g.	Silence	-

Figure 3-4: Structure of the binary decision tree

3.2.2 The Pattern Classification Machine

With such a classification structure, a binary decision is made at each of the nodes. To improve the tractability and formalism of the classification procedure, the same sub-classifier structure is used at each of the nodes. Thus the only possible difference at different nodes is the feature vectors used by he sub-classifiers. The structure of the pattern classification machine u it each of the nodes is shown in Figure 3-5.

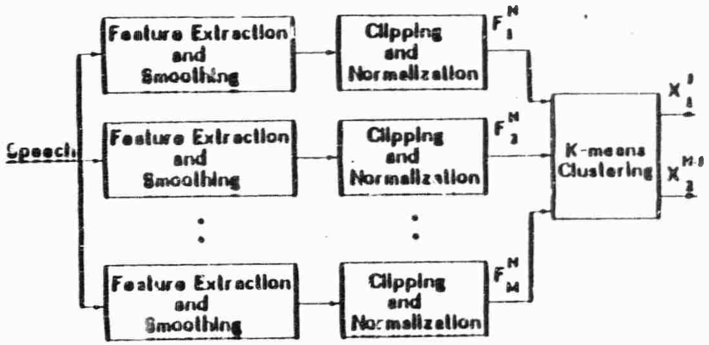

Speech
Feature Extraction and Smoothing
Feature Extraction and Smoothing
Feature Extraction and Smoothing
Clipping and Normalization
Clipping and Normalization
Clipping and Normalization
F^N_1
F^N_2
F^N_M
K-means Clustering
X^1_1
X^{M-1}_2

Figure 3-5: Block diagram of the pattern classification machine. Superscripts denote number of samples.

3.2.2.1 Feature Extraction and Smoothing

The speech signal is digitized at 16 KHz. Every 5 ms, an M-dimensional feature vector is obtained from the speech signal which has been multiplied by a 25.6 msec

Hamming window. Figure 3-6 shows the feature parameters used at each of the nodes shown in Figure 3-4. These feature parameters include energies at different frequency bands, normalized autocorrelation coefficient at unit sample delay [R(1)/R(0)], root mean square energy of the waveform [RMS], ratio of energies at different frequency bands [ER], first moments of spectral energy [COG], and zero crossing rate [ZCR]. All these feature parameters are selected based on acoustic-phonetic knowledge of the broad phonetic classes in American English. The feature parameters are smoothed by using a seven point median smoothing algorithm [Rabiner 75].

Node	Features
1	R(1)/R(0) RMS ZCR ER (125-750 Hz to 5000-8000 Hz) ER (750-2500 Hz to 5000-8000 Hz)
2	ZCR Energy (750 - 4000 Hz) Energy (3400 - 5000 Hz) Energy (5000 - 8000 Hz)
3	ZCR Energy (0 - 8000 Hz) Energy (750 - 4000 Hz) Energy (3400 - 8000 Hz) Energy (5000 - 8000 Hz)
4	ZCR ER (750-4000 Hz to 5000-8000 Hz)
5	RMS Energy (750 - 3000 Hz) COG (0 - 3000 Hz) ER (125-500 Hz to 750-3000 Hz)
6	Energy (0 - 8000 Hz) Energy (750 - 4000 Hz) Energy (1850 - 3000 Hz) Energy (3400 - 8000 Hz)

Figure 3-6: The features used at the nodes of the classification tree.

3.2.2.2 Feature Clipping and Normalization

The smoothed feature parameters are then clipped and normalized. Clipping emphasizes the portions of the feature parameters where boundaries are likely to occur. Figure 3-7 shows the zero crossing rate contour for a vowel-fricative transition. The vertical dashed line is drawn at the boundary between the phonetic events by an experienced transcriber. As can be seen in the Figure, the zero crossing rate climbs up to an extremely high level just a few ms. after the boundary. In fact, the zero crossing rate is so high that an obstruent is certain to be there. Thus a level of certainty can be established and the feature parameter can be clipped to this level. A similar level of certainty can also be established in the vowel portion. The horizontal dashed lines in Figure 3-7 represent these levels of certainty. As can be seen in the Figure, the phonetic boundary is still located between these levels of certainty. Thus clipping triggers a strong evidence for a particular broad class. The phonetic boundaries can then be located by more sophisticated techniques as will be described later.

The clipped parameters are normalized to the same scale before they are combined for classification. This is important because after the clipping procedure, the actual values of the feature parameters are not as important as the relative values. The normalization is performed by means of a linear transformation

$$y = [x - c1] / [c2 - c1]$$

where x is the original feature value, y is the transformed value, c1 and c2 are the

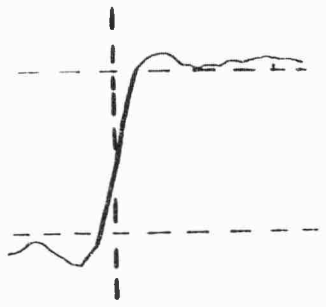


Figure 3-7: Zero crossing rate contour and levels of certainty for a vowel-fricative transition

levels of certainty. Together the clipping and normalization procedures effectively assign different weights to different feature parameters depending on how much the feature distributions of the two classes overlap. A feature parameter is reliable and robust if the clipping levels of certainty can be chosen close to each other. The closer they are, the more reliable the feature should be. However, the difference of these levels are inversely proportional to the transformed feature parameter, as seen from the above equation. Therefore, a more reliable feature is more heavily weighted.

3.2.2.3 A 2 means Clustering Algorithm

Every 5 ms, an M-dimensional feature vector is obtained. All the feature vectors

for a given utterance constitute samples in the feature space. A binary decision is then made in the M-dimensional feature space by using K-means clustering, with a Euclidean distance metric [Tou and Gonzalez 74]. However, the algorithm can become more powerful with the proper utilization of speech knowledge.

For illustrative purpose, Figure 3-8 shows a 2-dimensional feature space with a number of feature vectors to be classified into two classes. The circles denote samples from the sonorant-like segments of an utterance and the X's denote samples from the obstruent-like segments of the same utterance. The two feature parameters are zero crossing rate and normalized autocorrelation at unit sample delay. However, the identity of these circles and X's are unknown to the clustering algorithm. In the ideal case, the algorithm is expected to classify all the open circles into the sonorant class and all the X's into the obstruent class.

The location of the cluster centroids and the speed of convergence for a clustering algorithm depend on the choice of the initial centroids, the number of clusters, and the geometrical distribution of the data. Besides those mentioned in a previous section, the use of a binary classifier has a few more advantages in this particular system structure. With 2 clusters, the clustering algorithm always converges. Furthermore, by applying acoustic-phonetic knowledge, the initial centroids can also be selected at the appropriate extrema of the feature space.

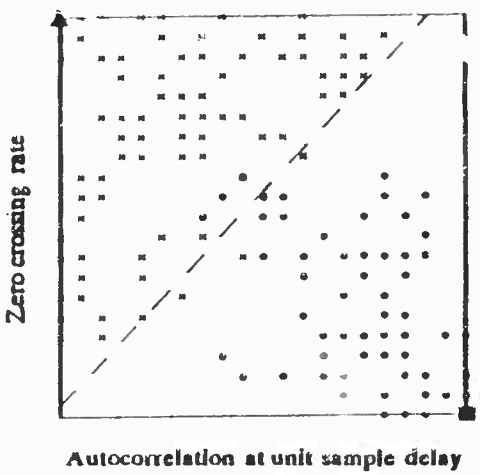


Figure 3-8: A 2-dimensional feature space with the initial centroids and the feature vectors to be classified.

3.2.2.4 Initial Cluster Centroids

The K-means clustering algorithm updates the centroids every iteration until it converges to a local minimum. As mentioned earlier, these cluster centroids depend on the locations of the initial centroids. In most applications of the clustering algorithm, the initial centroid locations are selected arbitrarily. In our situation, however, these initial centroids can be selected intelligently, since the general properties of the samples in the M-dimensional space are known. For example,

since the samples from sonorant-like regions of the utterance are expected to have high autocorrelation coefficients and low zero crossing rate, the initial sonorant centroid can be reasonably located at the proper extremum of the feature space, as shown by the filled square in Figure 3-8. The samples from the obstruent-like regions of the utterance are also expected to have low autocorrelation and high zero crossing rate. Therefore, the initial obstruent centroid can also be located at the other extremum of the feature space, as shown by the filled triangle in the Figure. After each iteration, the cluster centroids migrate towards each other until they come to a local minimum. A decision boundary can then be determined as shown by the dashed line in the Figure. All the samples of the same cluster are then assigned to the same class. In this case, three samples of the sonorant-like segments and three samples of the obstruent-like segments are misclassified to the wrong classes. These samples are very likely located at the transitions between the two types of segments.

With the proper selection of the initial centroid locations, the algorithm can converge to the desired minimum in about 5 to 7 iterations. However, the algorithm may converge to an undesirable local minimum and have the samples classified into the wrong class if the initial centroid locations are selected arbitrarily. Figure 3-9 shows the same 2-dimensional feature space. If the initial centroids are selected at the extreme marked by the triangles, the clustering algorithm with Euclidean distance metric may converge and draw the decision boundary as shown by the dashed line in the Figure. This is obviously a logic error, but it can be avoided by intelligent initial centroids.

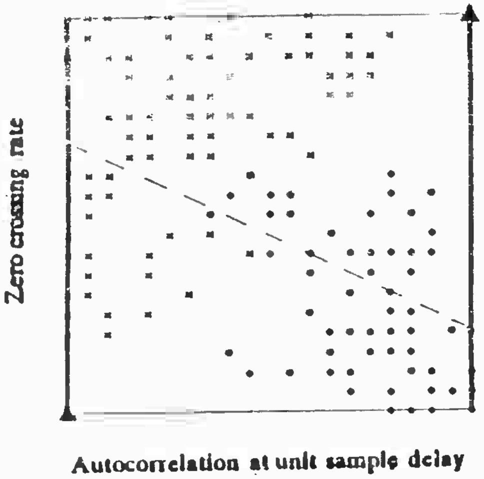


Figure 3-9: An example showing how the clustering algorithm fails with an arbitrary selection of initial cluster centroid locations.

By using such a classification structure, all the samples being clustered come from the same utterance and, therefore, from the same speaker and the same recording environment at about the same time. This is a very desirable feature because the iterative procedure can adapt to the speaker characteristics at the recording time. It can also adapt to the recording environment and thus is more immune to

background noise. One potential disadvantage of such a classification technique is that the algorithm forces the input data into two classes. This is particularly a serious problem when all the input data should all be classified into one broad class category. However, this problem can be solved by artificially introducing a few fictitious feature vectors at the initial centroids. They are then used along with the actual feature vectors obtained from the utterance as input to the clustering algorithm. Thus the actual feature vectors are not forced into two classes. After the iterative algorithm converges, the fictitious feature vectors are disregarded.

3.2.2.5 Levels of Certainty for the Feature Parameters

Determining the levels of certainty of the feature parameters is important. On one hand, they should be selected as close as possible so that the feature parameter can have an appropriate weight. On the other hand, they should be enough apart so that the phonetic boundaries fall within the transitional region between the two levels of certainty. In order to determine these levels reliably, statistics of the feature parameters are obtained from different sounds of American English. Approximately 30 seconds of manual aligned speech used in the feasibility study is randomly chosen and used for training. The levels of certainty can then be determined by predetermined levels of significance.

Figure 3-10 shows the distribution of a feature parameter for two different classes of speech frames. From the Figure, it can be seen that the levels of certainty should be chosen at points A and B. If a speech frame has a feature measurement higher than point A or lower than point B, then it can be classified into the appropriate class with high confidence. However, it should be noted that this clipping procedure is not the final decision about the destiny of the frame. All the feature vectors from the utterance, clipped and not clipped, are input of the iterative clustering process. If one feature parameter of a speech frame is clipped to one class but the others are not, the frame may still be classified to the other class by the iterative process.

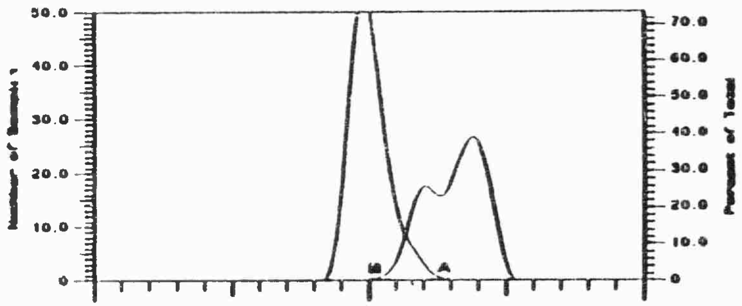


Figure 3-10: A feature parameter distribution for two classes of data

In practice, the clipping levels are not chosen at points A and B. This is because some feature distributions have a small number of outliers and thus resulting in long-tailed distributions. Thus a set of levels of significance is pre-determined.

Figure 3-11 shows the level of confidence used at each of the decision nodes. These levels range from 94% to 99.9% depending upon the specific classification. A lower level of significance is used for more subtle phonetic transitions, since the manual alignment for these transitions are also less robust.

Node	Confidence level (%)
1	99.9
2	94.0
3	99.0
4	97.0
5	97.0
6	97.0

Figure 3-11: Levels of confidence for clipping

3.2.3 An Example Illustrating the Performance of the Initial Classifier

Figure 3-12 illustrates the results of the initial classifier. The speech waveform is shown at the top of the Figure. The hand transcription is shown above the spectrogram. The labels of each of the segments are shown in the row labeled "Segmentation Output". For example, the classifier found an obstruent for the word initial /g/, an obstruent for the phoneme /ʃ/ , whereas all segments in between were merged into one long sonorant region. It can be seen that the initial segmentation compares favorably well with the hand transcription. Since the hand alignment is performed on the speech waveform which is sampled at 16 KHz, and the automatic alignment is performed on features computed 200 times a second, the hand transcription has a 80 times better resolution than the automatic procedure.

3.2.4 Comparison with Clustering Based on Linear Predictive Coefficients

As has been described earlier, the goal of the iterative clustering algorithm is to divide the input speech signal into meaningful broad acoustic-phonetic classes. In order to do so, the clustering process must be performed in an appropriate M-dimensional feature space. Thus it is essential to select the features judiciously based on acoustic-phonetic knowledge. At the present time, most of the suggested clustering algorithm in the speech processing literature uses straight-forward representation of the speech signal such as linear predictor coefficients [Rabiner and Schafer 78]. While this approach may be appropriate for generating clusters with some optimality criterion, the resulting clusters may not form any significant acoustic-phonetic classes.

In order to have a more direct contrast between the use of straight-forward spectral representation and the judicious choice of the features based on speech

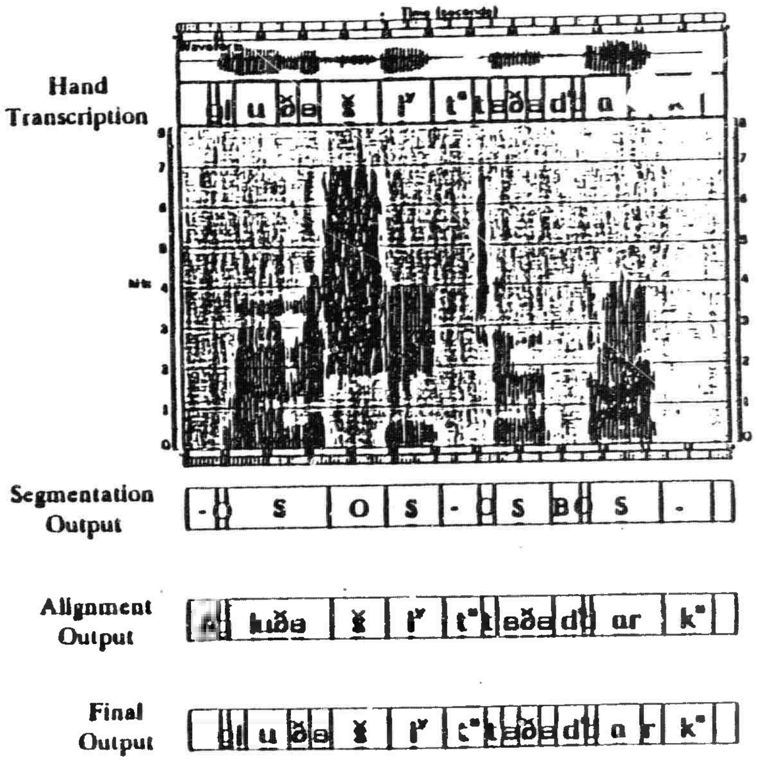


Figure 3-12: An example for the phrase "Glue the sheet to the.."

knowledge, the pattern classifier described earlier is slightly modified. Linear predictor coefficients are used instead of specific feature parameters. However, since the significance of each of the coefficients is not understood, the smoothing, clipping and normalization procedures were not used.

Every 5 ms, a 19-pole linear prediction analysis is performed on the Input speech signal, which is sampled at 16 KHz. Sixteen of the nineteen poles are used to represent the resonances of the vocal tract. Two are used to represent the glottal source, and one is used to represent radiation from the lips. The K-means clustering algorithm is then performed, with Itakura distance metric [Itakura 75]. At the end of each iteration, two cluster centroids are computed by averaging the correlation coefficients of the speech frames that are in the same cluster. This amounts to averaging the magnitude of the Fourier Transform. The corresponding linear prediction coefficients are then computed from the averaged coefficients, and the process repeats until the algorithm converges.

Figure 3-13 shows a comparison between the classification results based on linear predictor coefficients and selected features. For this example, the speech signal is to be classified into two classes. The top trace shows the result based on judicious choice of feature parameters. A high level corresponds to "sonorants" of the utterance, and a low level corresponds to "n[illegible] sonorants" of the utterance. The second trace shows the result based on linear predictive coefficients. A comparison of these two traces shows that they agree very much with each other.

Figure 3-14 shows the comparison with a different utterance. It can be seen that the binary classification based on [illegible]lected features still works very well. However, the classification based on linear predictive coefficients is not acceptable. A comparison with the spectrogram shows that it classifies the phoneme /8/ with the

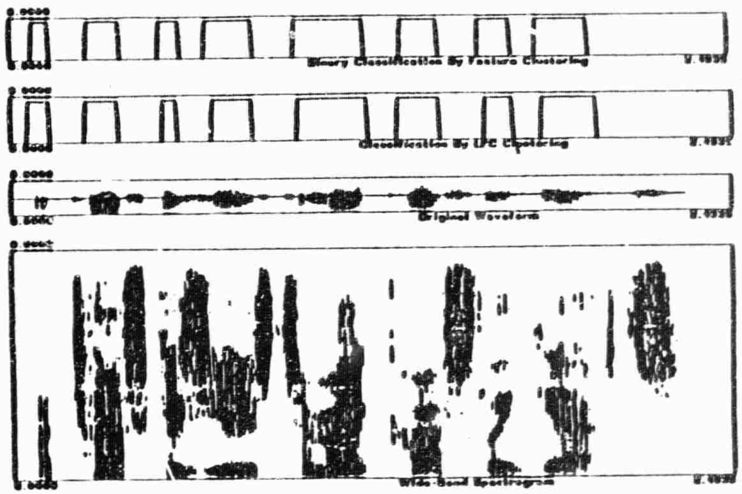


Figure 3-13: A performance comparison: The classification based on LPC compares well with the classification based on pre-selected features.

adjacent sonorants into the same class. Furthermore, it also gives occasional spurious segments.

As another example, Figure 3-15 shows another comparison with the same utterance. The speech signal is further classified into a total of 4 classes. The classification based on pre-determined features remains superior to the one based on LPC. In the latter, while the phoneme /8/ and its adjacent sonorants are now classified into different classes, the final result is still unacceptable due to its

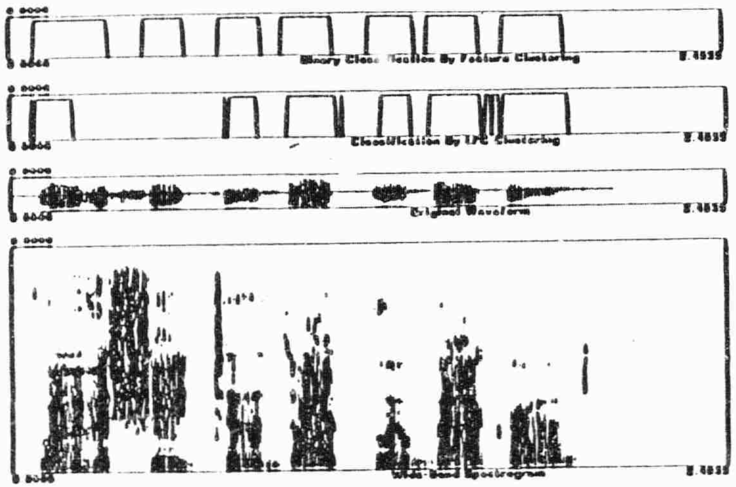


Figure 3-14: A performance comparison: The classification based on pre-selected features is superior to the classification based on LPC.

jaggedness. Furthermore, the 4 classes do not have any significant acoustic-phonetic correspondence.

All these experiments and the high performance of FEATURE discussed in Chapter 2 are consistent with the fact that the proper utilization of acoustic-phonetic knowledge is very important to speech related systems. Although the amount of speech knowledge incorporated is not very much, the system performance can improve significantly.

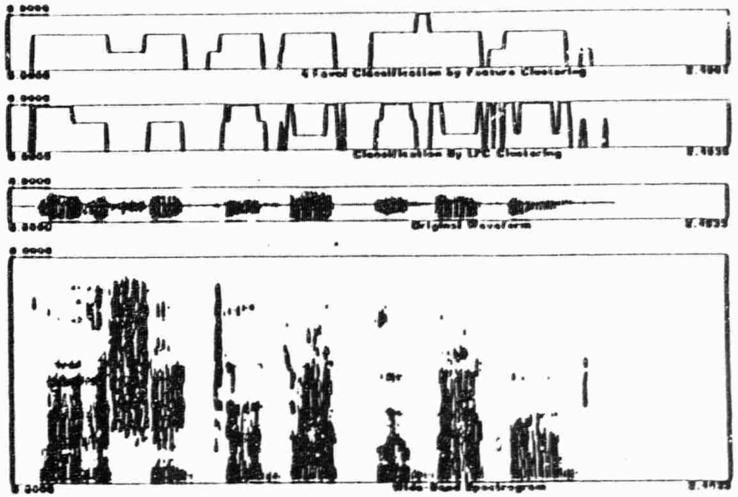


Figure 3-16: A performance comparison: The classification based on LPC becomes jagged as the number of classes increases.

3.3 Alignment of Broad Class Segments with the Phonetic Transcriptions

The output of the initial classifier is a broad, but presumably robust description of the significant acoustic-phonetic events in the speech signal. In order to use this broad phonetic description as anchor points for more detailed analysis, the broad representation must now be aligned with the phonetic transcription. This can be viewed as a path finding problem in which one attempts to match the two symbolic descriptions of the utterance. Since the two sequences of symbols carry linguistic content and acoustic realization of the utterance, acoustic-phonetic heuristics can be utilized to search for the best path.

3.3.1 Different Possible Ways to Match the Symbols

There are many ways these symbols can be matched. The acoustic to phonetic symbol mapping can be (1) one-to-one mapping, (2) one-to-many mapping, (3) many-to-one mapping, or (4) many-to-many mapping. The first mapping associates an acoustic segment or symbol with a phonetic event or symbol. However, due to the acoustic ambiguity in the speech signal and the fact that broad categories are used, a phonetic event may be mapped to different acoustic segment in different context. The second mapping corresponds to "deletion error" in the initial segmentation. This is also expected to happen often since the segmentation is intended to be performed at the broad level. As can be seen in Figure 3-16, although the word "lie" is actually a liquid followed by a vowel, its acoustic realization is very much like a long sonorant and there is no distinct acoustic cues for the transition. In this case, there is a one-to-two mapping. A one-to-many mapping also indicates that further processing is necessary to divide the segment into more detailed phonetic events. The third mapping corresponds to "insertion error" in the initial segmentation. This kind of mapping is not expected to happen often. In fact, it is intended to bias the initial segmentation towards "deletion error" as opposed to "insertion error". Although no further segmentation is needed, this kind of

mapping can signal a possible error. On the other hand, this mapping is expected to happen to certain phonetic events more frequent than the others. For example, the phonetic event, /f/, is quite frequently realized as a sequence of silence followed by a weak turbulence noise due to its labial articulation. The fourth mapping signals a great deal of ambiguity in the region and thus further intensive processing is needed within the entire segment.

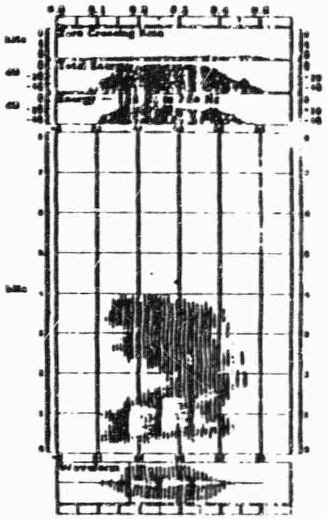


Figure 3-16: A spectrogram showing the similar acoustic realizations of the two phonemes in the word, "lie".

3.3.2 A Heuristic Path Finding Approach

For a sentence of about 2 seconds long, there are approximately 30 phonetic events. With a branching factor of 5, the number of possible paths is approximately 5^{30}, or 9×10^{20}. Thus an exhaustive search is impractical. Even if an exhaustive search were done, there would still be a great deal of ambiguity in determining the correct and unique path to match the two sequences of symbols.

There are at least 2 possible ways to realistically match the broad acoustic classes with the phonetic events. One is to utilize acoustic-phonetic knowledge in the form of production rules to constrain the search space to only one allowable path. This is an attractive approach because of the rich acoustic-phonetic information in the search space. However, due to the acoustic ambiguity in the speech signal, it is difficult to have rules to take care of all the different ambiguities in different environment. Another approach is to associate cost to each matching of the symbols. However, some kind of pruning is necessary in order to make this approach computationally efficient. The pruning must also be performed carefully enough so that the right path will not be eliminated.

Neither of the two above methods alone can effectively find the right path to match the two sequences of symbols. A method that combines the above two methods have been chosen. General heuristic rules can then be written to constrain the path. Besides constraining the number of permissible paths, the rules can also resolve some of the ambiguities. With these rules, a path finding algorithm can then

be used to search for the minimal cost path to match the broad acoustic segments with the phonetic transcription.

Figure 3-17 shows the search space for the alignment of the same utterance as shown previously. The horizontal dimension represents the output of the broad classifier, while the vertical dimension represents the actual phonetic transcription. The problem now is to start from the upper left corner of the search space, and traverse through the state space until the lower right corner is reached. In other words, beginning from the start node at the upper left corner, successor nodes are expanded. Pointers are set up from each successor back to its parent node. These pointers indicate a path back to the start node when the goal node is finally found. The successor nodes are checked to see if they are the goal node. If the goal node has not yet been found, the process of expanding nodes continues. When a goal node is found, the pointers are traced back to the start node to produce a solution path.

3.3.3 Local Path Constraints

In order to have a practical and meaningful path, constraints to what nodes to expand must be applied. Figure 3-18 illustrates the local path constraints to expand a node. Different paths provide different ways to expand a node. Path 1 corresponds to skipping an acoustic segment. This allows one phonetic event to match with a sequence of acoustic segments. Path 3 corresponds to matching an

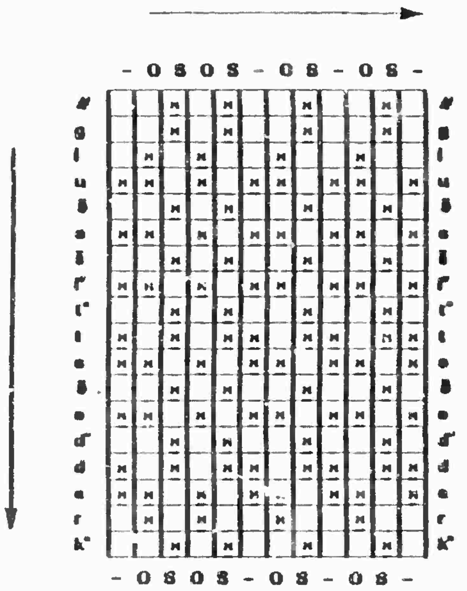


Figure 3-17: The alignment search space for the same phrase,
"Glue the sheet to the.. ".

acoustic symbol with a phonetic event. Path 5 corresponds to skipping a phonetic symbol. This allows multiple phonetic events to match with one broad acoustic segment. Paths 2 and 4, however, provide alternative ways to skip a phonetic or acoustic symbol respectively. For example, if a vowel is followed by a voiced stop, /b/, /d/, /g/, the stops are often very weak (if it exists) and thus the pattern

classifier might not be able to detect such a weak release. In this case, path 5 will allow the second symbol, the voiced stop, to be skipped. However, if the voiced stop is pre-vocalic, i.e. a voiced stop followed by a vowel, the first phonetic symbol should then be skipped. Path 4 provides a way to skip the pre-vocalic stop. Thus path 1 and path 2 provide two ways to skip an acoustic symbol, path 4 and path 5 provide two ways to skip a phonetic symbol. Path 3 provides a good match of the symbols.

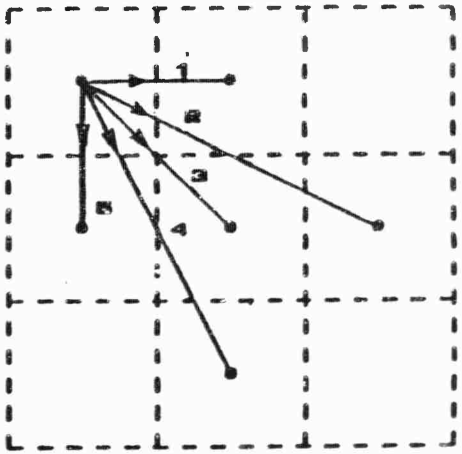


Figure 3-18: A diagram showing the local constraints to expand a node.

Figure 3-19 shows the tree representation for the search space problem. The start node is shown at the top of the tree. Each node can have 5 possible successors, as determined by the local path constraints. Associated with each node are its coordinates in the state-space. As can be seen in the Figure, there are more than one way a state can be reached from the start node. In fact, the number of ways a state can be reached increases exponentially as the state gets farther and farther away from the start node. Thus, in finding the best path, some kind of heuristics is necessary to make the finding feasible.

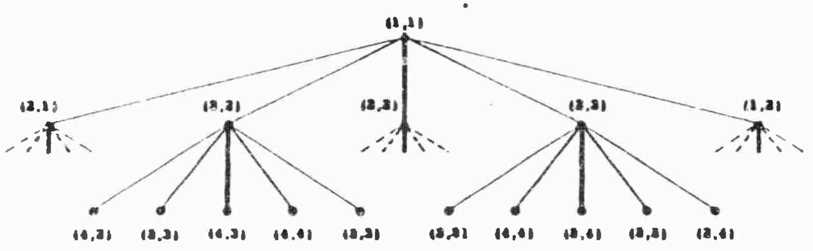


Figure 3-18: A search tree representation

3.3.4 Application of Acoustic-phonetic Heuristics

Acoustic-phonetic heuristics are utilized in two forms. One is to associate each node a cost or evaluation function. This function should be an estimate of the cost of a minimal cost path from the start node to the goal node constrained to go through that node. The purpose of using cost functions is for ranking those nodes that are candidates for expansion to determine which one is most likely to be on the

best path to the goal. By doing so, segmentation errors, such as segment deletion, or mislabeling produced by the initial classifier can also be handled. An algorithm can then be used in which that unexpanded node having the smallest cost is selected for expansion next.

Acoustic-phonetic heuristics are also utilized in the form of production rules. These rules can be used to terminate unreasonable paths before the goal node is reached. There are two kinds of rules. First, the path is not allowed to traverse through certain nodes, since this will produce implausible phonetic segments. These mismatches are stored as a set of context-independent rules. For example, a strong fricative is not allowed to match with a sonorant acoustic segment found by the initial classifier. Second, there is a set of contextual constraints that can terminate certain paths. For example, although a semi-vowel alone is allowed to match with a "S" or "B" segment, it is not allowed to match with a sequence of the two. Furthermore, durational constraints can also terminate unreasonable paths. For example, the minimum duration of an "O" segment to match the phoneme, /s/, is 0.050 seconds. In addition to reducing the amount of computation, the termination procedures can also disambiguate some of the uncertainties during the path finding process. There are altogether approximately fifty such rules.

3.3.5 Different Path Finding Strategies

Two path finding strategies have been investigated. One is based on the principle of dynamic programming and the other is based on branch and bound [Winston 84]. The principle of dynai ic programming is as follows. If a state can be reached from the start node in differen' ways, t' en the state is only associated with the partial path that has the smallest cost. All the other partial pat' s to that state would be disregarded. Note that the concept of dynamic programming here is not quite the ame as that in the literature of isolated word recognition where dynamic time-warping is performed on a frame-by-frame basis. The advantage of using dynamic programming is that the number of nodes expanded can be reduced tremendously since each state is associated with only one partial path from the start node. However, there is also a disadvantage.

As has been mentioned earlier, production rules are used to guide the alignment of the symbols and erminate unreasonable paths. These rules may be context dependent and may also depend on the previous nodes in the path. Thus the partial path from the start node to an intermediate node does influence the ch ice of paths for traveling from the the same intermediate node to the goal node. Consequently, the principle of dynamic programming can no longer find the complete optimal path. Figure 3-20 shows how the principle of dynamic programming with path constraints misleads the path. Note that no A can be reached from the start node via 2 different paths. However, each one is associated with a different cost. The use of dynamic programming will discard the one that has a higher cost of 7 and only keep the one that has a lower cost of 5. Unfortunately, the one with the lower cost can never reach the goal node, G.

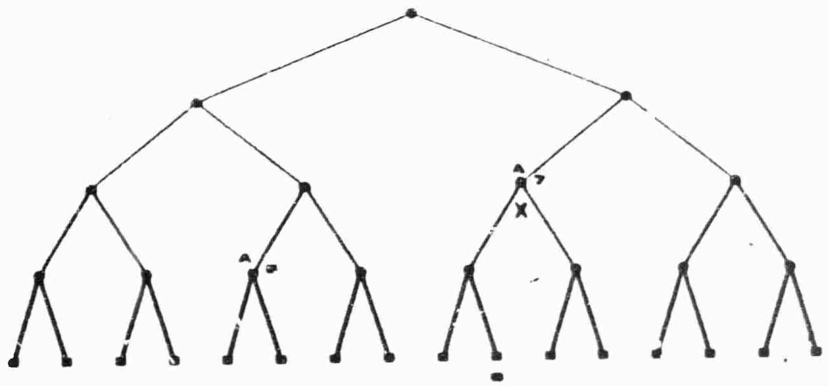


Figure 3-20: A search tree showing how the principle of dynamic programming leads to a fatal search of the goal node.

To avoid this hill-climbing disadvantage, a branch and bound search is conducted instead. This search algorithm sorts all the unexpanded node and expands the one with the lowest cost. No pruning is performed. Thus, in Figure 3-20, that node A with the lower cost will be expanded first. However, the other node A with a higher cost will also have a chance to expand when it gets to become the unexpanded node with the lowest cost.

3.3.6 Examples

The alignment mechanism of the same utterance is shown in Figure 3-21. The context-independent mismatches are marked by "X" in the Figure. For example, the phoneme /g/ is not allowed to match with a sonorant. As an example of the contextual constraints, the second schwa in the utterance is not allowed to match with a sequence of sonorant and voice-bar. As another example, although the phoneme /ʒ/ is an obstruent, it is not allowed to match with the second obstruent due to a durational constraint. By applying these rules and the cost functions, the final path is found, as illustrated by the filled circles in the Figure. For illustrate purpose, the open circles denote all the states that have been expanded during the path finding process. The application of acoustic-phonetic knowledge in the form of rules can greatly reduce the number of permissible paths.

The row labeled "Alignment Output" of Figure 3-12 illustrates the results of the alignment obtained from the complete path shown in the previous Figure. It can be seen that in some cases, there is only one phonetic event in a segment. In the other cases, there can be as many as four phonetic events. Furthermore, a comparison with the hand transcription shows that all the phonetic events are mapped to the right broad class segment.

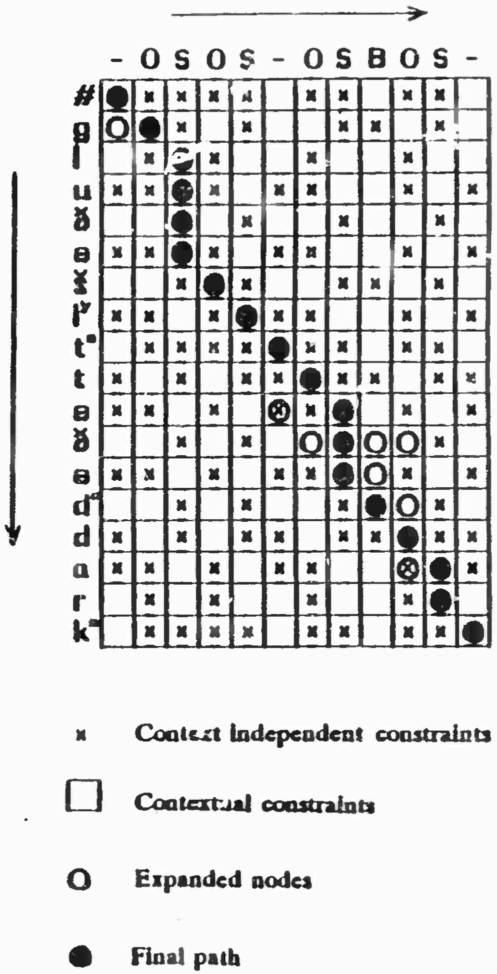


Figure 3-21: An example showing the final path found and the nodes expanded.

Figure 3-22 shows an example in which the application of dynamic programming results in a wrong alignment. The top trace shows the result after applying the principle of dynamic programming, whereas the second trace shows the result after applying the principle of branch and bound. As can be seen in the Figure, application of dynamic programming results in misalignment of the phonetic events that correspond to "slid on the" in the utterance.

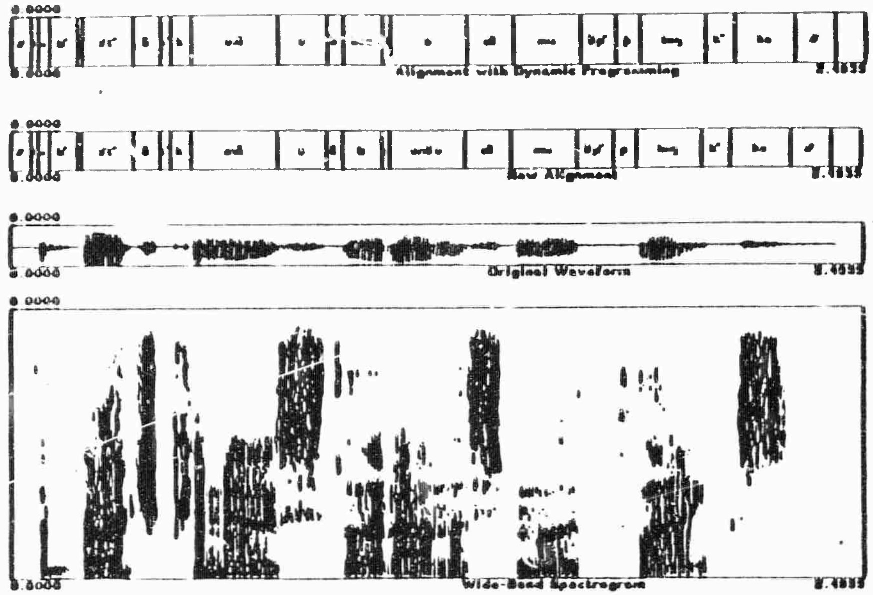


Figure 3-22: An example showing differences between the performance of dynamic programming and branch-bound algorithm.

3.4 Context-dependent Segmentation

The path finding algorithm described divides the speech signal into a sequence of segments. Each segment is mapped into one or more phonetic symbols or events. No further processing is needed if there are one or more segments mapped into one phonetic event. For those segments which correspond to 2 or more phonetic events, further segmentation of the speech signal is necessary. These mappings provide anchor point which, in turn, provide information about duration of a segment, the number of phonetic events in a segment, and the exact nature of the phonetic events in a segment. This knowledge makes classification of the speech signal into more detailed segments much easier. Thus we can choose specific processing for specific environment.

There are, in general, two kinds of processing to mark a new boundary between the anchor points. One is to apply heuristic and consistent rules to transitions which are gradual and not marked by any distinct acoustic cues. For instance, pre- and post-vocalic liquids right next to certain vowels are assumed to have a duration that constitutes one-third of the syllable nucleus. If a silence segment is matched with a sequence of two stop closures, then the two closures are assumed to have equal duration.

For phonetic transitions that are more pronounced, appropriate algorithm and features based on the phonetic environment can then be selected. Since the exact phonetic events are known, the segmentation can be much more specific. There are two major algorithms. One algorithm is for detecting and locating onset or offset of some pre-selected feature parameters. For example, to locate a boundary between a vowel and a nasal, we can use the same classification machine used in the initial segmentation. By using energies at the low frequency bands, the pattern classifier can classify the whole segment into 2 regions, one having higher energies and the other one having lower energies. The other algorithm is for detecting and locating a dip or a spike of a feature parameter. For example, if a "S" segment is matched with a sequence of vowel, voiced weak fricative, vowel, the intervocalic voiced weak fricative is expecred to have a slight dip in low frequency energy. The dip can be located by using a Gaussian edge detector:

$$G(r) = \left(1 - \frac{r^2}{\sigma^2}\right) e^{-\frac{r^2}{2\sigma^2}}$$

where r denotes the feature parameter. Figure 3-23 shows the above function. The total area under the curve is equal to zero with 50% of the area under the main lobe. The zero crossings, however, are only one standard deviation away from the center of the main lobe. This narrow and high main lobe makes the filter particularly suitable for detecting dips or spikes. With proper selections of the features, the edge detector can locate most of the dips and spikes, the edges of which are at the zero crossings of the filtered feature parameter.

These two kinds of processing, namely, rule-based and feature-based

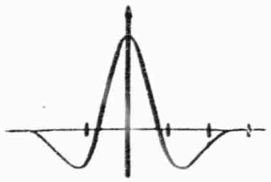


Figure 3-23: A Gaussian edge detector

segmentation can be applied recursively to locate the detailed phonetic events. The row labeled "Final Output" of Figure 3-12 shows the output of the knowledge-based segmentation for the same utterance shown earlier. From the output at the second stage, it is known that the intervocalic voiced fricative, /ð/, is embedded in the first sonorant segment. By using the Gaussian filter and the energy contour from 500 Hz to 3000 Hz, this /ð/ can be located as shown. The phoneme, /l/ can then be delineated from the vowel using the rule that it constitutes one-third of the syllable nucleus. The same edge detection can be used to locate the /ð/ embedded in the third sonorant segment. The phoneme, /r/, also constitutes one-third of the last sonorant segment. All broad segments with only one phonetic event remains unchanged. In this example, all the phonetic events have been located. There are altogether approximately fifteen to twenty specific algorithms in the third stage of the system.

3.5 Chapter Summary

- The implemented automatic alignment system has three major components: an initial broad classifier, a heuristic path finding algorithm, and a knowledge-based segmenter.

- The initial classifier is structured as a sequence of identical binary classifiers. This non-parametric classifier makes no assumption about the distributions of the featurc parameters and needs no intensive training. By incorporating specific speech knowledge, the classifiers can adapt to different speaker characteristics.

- By matching the output of the initial broad classifier with the phonetic transcriptions, the heuristic path finding algorithm provides "islands of reliability" for more context-dependent segmentation.

- By utilizing specific algorithms and featur s based on the immediate phonetic contexts, the knowledge-based segmenter can locate more detailed phonetic events in the utterance.

Chapter Four

Performance and Evaluations

The system was evaluated using the same set of utterances in the feasibility study discussed in Chapter 2. There are 40 distinct sentences, randomly chosen from the Harvard List. Five talkers, three male and two female, each read twenty sentences, for a total of one hundred (100) tokens. The corpus contains approximately 4 minutes of speech material and twenty seven hundred (2700) phonetic events. All sentences were hand transcribed by an experienced acoustic phonetician.

Figure 4-1 shows the statistics on the number of phonetic events in a segment produced by the initial classifier. Approximately 80% of the time there is only one phcnetic event in a segment, whereas 17% of the time there are two phonetic events in a segment. These results compare favorably with the ideal segmentation described in Chapter 2. For the ideal segmentation, the corresponding figures are 85% and 12% respectively. A closer comparison of this Figure with Figure 2-5 shows that there are more "S" segments in the actual segmentation than in the ideal segmentation. This is because the initial classifier can sometimes segment a sequence of vowel-liquid-vowel into three segments, whereas the ideal segmentation always gives one long segment. There are also more "-" segments because some of the weak fricatives and voiced stop closures are classified into this category.

SEGMENT LABEL	1	2	3	4	5	TOTAL
S	497 (63%)	226 (29%)	35 (4%)	26 (3%)	6 (1%)	790
O	597 (86%)	79 (11%)	18 (3%)	2 -	1 -	697
B	113 (81%)	22 (16%)	4 (3%)	0 -	0 -	139
-	555 (89%)	62 (10%)	4 (1%)	2 -	0 -	623
D	91 (88%)	13 (12%)	0 -	0 -	0 -	104
TOTAL	1853 (79%)	402 (17%)	61 (3%)	30 (1%)	7 -	2353

Figure 4-1: Statistics on number of phonetic events in one segment produced by the initial broad classifier.

Figure 4-2 summaries the confusion statistics between the broad class segments and the phonetic classes. Over 90% of most of the phonetic classes are classified into one or two acoustic labels. Vowels and strong fricatives are most robust and over 97% of them are classified into the "S" and "O" segments respectively. Nasals, however, have their acoustic realizations more dependent on their phonetic context. Consequently, the segmentation and labeling of these phonetic classes are not as consistent as the other classes.

Phonetic class	S	O	B	-	D	Total
Vowel	480 (97%)	7 (1%)	8 (1%)	5 (1%)	2 -	502
Nasal	11 (11%)	1 -	55 (57%)	9 (9%)	21 (23%)	97
Liquid	5 (8%)	0 -	29 (47%)	0 -	28 (45%)	52
Strong fricative/ affricate	0 -	201 (98%)	0 -	3 (1%)	2 (1%)	206
Weak fricative	1 -	106 (68%)	6 (4%)	36 (23%)	8 (5%)	157
Stop release	0 -	247 (99%)	0 -	4 (1%)	0 -	251
Stop closure/ silence	0 -	26 (5%)	15 (3%)	494 (91%)	6 (1%)	541

Figure 4-2: Confusion statistics between the broad phonetic classes and the segmentation labels.

Figure 4-3 shows the percentage of phonetic events located after two stages of processing. A phonetic event is located when there is a one-to-one correspondence between it and an acoustic segment. We see that approximately 80% of the phonetic events have been located after the alignment procedure. This number increased to 97% after knowledge-based segmentation. In other words, instead of requiring an expert to time-align all the boundaries, only 3% of the work needs to be done.

The reliability of the boundaries found by the system with those found by an

Number of phonetic events in 1 segment	Knowledge-based Dynamic Programming	Knowledge-based Segmentation
1	81%	97%
2	16%	2%
3 or more	3%	1%

Figure 4-3: Statistics on number of phonetic events in one segment after two different stages of processing.

experienced transcriber has also been compared. This is done by computing the absolute difference between the two sets of boundaries. Curve A in Figure 4-4 shows the cumulative distribution of the offsets between boundaries determined by the automatic alignment system and those by the acoustic phonetician. We see that approximately 75% of the boundaries are within 10 msec of each other, and over 90% of the boundaries are within 20 msec.

In order to measure inter-transcriber agreement, five of the one hundred (100) sentences, were manually labeled by a second acoustic phonetician. The cumulative distribution of the boundary offset between the two phoneticians is shown by curve B . Since it is difficult to say exactly where a boundary should be, this curve gives

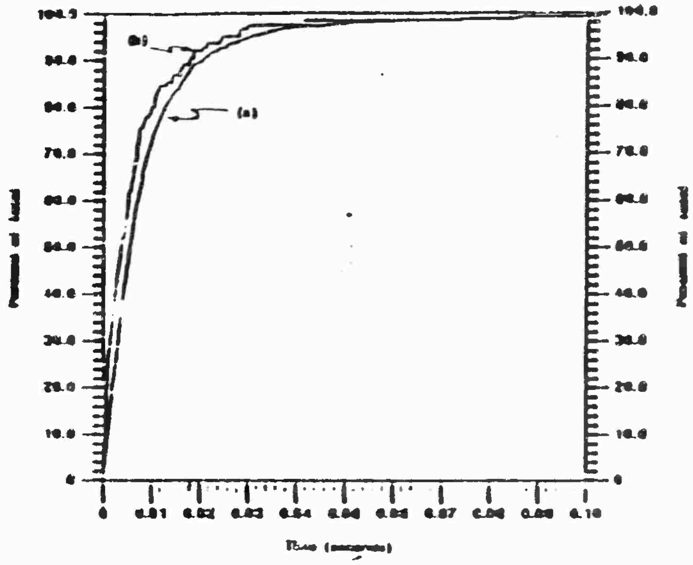


Figure 4-4: Cumulative distributions of the boundary offset.

an indication of the performance of the alignment system. We see that the system-transcriber differences are similar in magnitude to the inter-transcriber differences.

As has been discussed in Chapter 1, there is a continuum of reliability of the acoustic landmarks between phonetic events. Some landmarks are evidenced by distinct acoustic cues, whereas others are more subtle and even hand alignment cannot locate the landmarks reliably. Thus it is important to evaluate the automatic system by comparing the reliable landmarks found by the experienced transcriber with those found by the system. Figure 4-5 shows the mapping of the phonetic transitions into "hard" and "soft" transitions. For example, the transition from a strong fricative to a vowel is considered to be robust and "hard", whereas the transition from a vowel to a liquid is considered to be subtle and "soft". Transitions between two segments of the same class are always considered "soft", as can be seen at the diagonal of the Table.

	Vowel	Nasal	Liquid	Strong fricative/ affricate	Weak fricative	Stop release	Stop closure
Vowel	S	S	S	H	H	H	H
Nasal		S	S	H	H	H	H
Liquid			S	H	H	H	H
Strong fricative/ affricate				S	S	S	H
Weak fricative					S	S	S
Stop release						S	H
Stop closure							S

Figure 4-5: A table showing mapping of the phonetic transitions into "hard" and "soft" categories.

Figure 4-6 shows the cumulative distributions of these two kinds of boundary offsets relative to manual alignment. It can be seen that approximately 80% of the "hard" boundaries are within 10 msec from the manual alignment, whereas 80% of the "soft" boundaries are within 20 msec from the manual alignment. This performance difference can also be due to the inconsistent manual alignment of the "soft" boundaries. For example, while it is difficult for manual alignment to locate exactly one-third of the syllable nucleus, the automatic system can do so much more reliably.

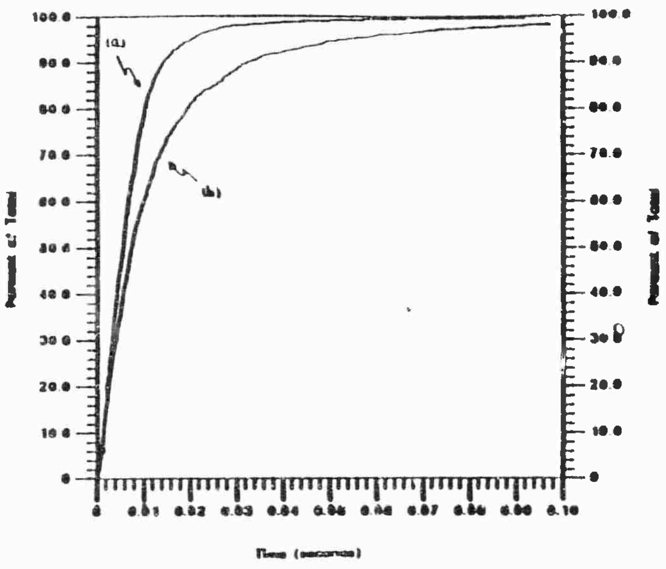


Figure 4-8: Cumulative distributions of the "hard" and "soft" boundary offset.

Chapter Five

Conclusions

5.1 Summary

In this thesis, we have proposed, designed and implemented a system for automatic alignment of phonetic transcriptions with continuous speech. Our motivation is threefold. First, it enables speech researchers to establish a large database and study the properties and characteristics of speech sounds in different phonetic contexts. These studies, in turn, can lead to a better model for speech production, as well as better rules for speech recognition and synthesis. Second, it can avoid manual alignment, which is tedious and inconsistent. Third, it can serve as a testbed for phonetic recognition.

There are three basic components in the designed system. First, the speech signal is segmented into broad classes using a non-parametric pattern classifier. This classifier has no knowledge about the phonetic transcriptions, makes no assumptions about the distributions of the feature parameters and does not need any intensive training. Second, a path finding algorithm, augmented with speech knowledge, then aligns the broad classes with the phonetic transcriptions. In addition to matching the phonetic transcriptions with the output of the initial classifier, the alignment can also "correct" segmentation errors. By doing alignment at the phonetic level, the system can often tolerate inter- and intra-speaker variability. These aligned broad classes also provide "islands of reliability", and establish phonetic contexts for further processing. Third, by proper utilization of speech knowledge, specific algorithms and features can be selected to align the more detailed phonetic events. Acoustic-phonetic knowledge is used extensively throughout the system.

In summary, the original three goals of the thesis have been achieved. We are encouraged by the preliminary results, and are hopeful that the system will play a major role in establishing a large database for speech research. Furthermore, the automatic alignment system also proposes a possible approach to phonetic recognition. After establishing the major and robust acoustic-phonetic events in the speech signal through use of broad classification and alignment, the specific phonetic events can be recognized by applying more specific speech knowledge.

5.2 Suggestions for Future Research

Some aspects of this thesis work are interesting for further research or improvement. In particular, there are four major areas that may be worth pursuing.

First, the cost functions used in the alignment procedure are now determined heuristically. A more robust method is to obtain the cost functions statistically. This requires a larger set of training data and observe how the acoustic segments match with the phonetic transcription.

Second, since the initial classification does not make use of the phonetic transcriptions, it is reasonable to expect the boundaries so found can be improved. After the alignment process, these boundaries can be further improved by applying specific features and algorithms.

Third, until now, the alignment system makes no attempts to locate the acoustic landmark between two phonetic events of the same class, such as the vowel-vowel transition. Since the system is developed in such a way that continuously more algorithms can be easily added on, these kinds of acoustic landmarks can be located by applying more specific speech knowledge.

Finally, although the system can greatly reduce the amount of time in establishing a speech database, it still requires an experienced acoustic phonetician to provide the phonetic transcriptions. An interesting area of research is to study how the phonetic transcriptions can be aligned with the orthographic transcriptions directly. Such study will require better understanding and applications of acoustic-phonetic knowledge, as well as more sophisticated pattern classification techniques and knowledge representation. This study, in turn, can also lead to better understanding of phonetic recognition.

Appendix A

Implementation and Human-machine Interface

A.1 Implementation

The automatic alignment system has been implemented on a Symbolics 3600 Lisp Machine, with a FPS100 array processor. Most of the feature parameters are computed at the array processor. Both the clustering and the path finding procedures are performed on the Lisp Machine. The entire system is built on SPIRE, a speech and phonetic interactive research environment developed at MIT.

Speed of the system is not a major concern because alignment of utterances can be submitted as batch jobs. Nevertheless, the system speed has been optimized to a certain extent. It now runs at approximately 30-40 times real-time.

A.2 Human-machine Interface

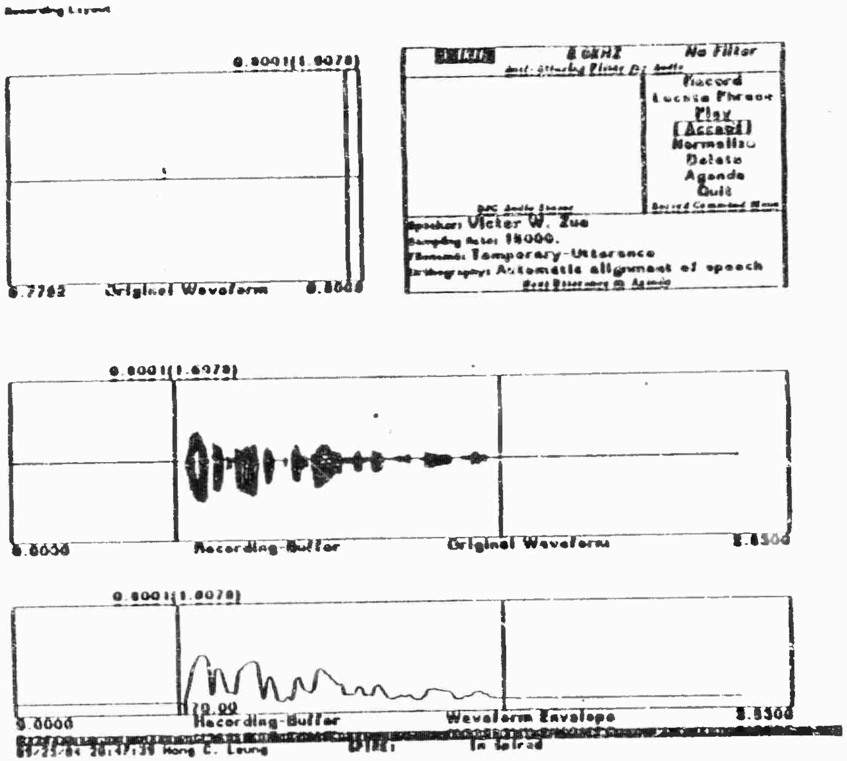
Before alignment of the phonetic transcriptions with the speech signal is performed, the speech signal must be recorded and digitized. The phonetic transcriptions must also be entered into the system by an experienced acoustic-phonetician. After the automatic alignment is performed, the output can be checked by the acoustic-phonetician and the acoustic landmarks can be adjusted, if

needed. This final adjustment is an optional process depending on the degree of resolution and accuracy required in a particular study.

Figure 5-1 shows a recording layout. In this layout, the speech utterance can be digitized from an audio cassette tape or directly from a microphone. An automatic end-point detector can then locate the utterance [Lamel 81]. The identity of the speaker and the orthographic transcription can also be entered into the computer.

Figure 5-2 shows a phonetic events layout. In this layout, the sequence of phonetic events can be entered into the computer by mousing the appropriate item on the menu. When necessary, portions of the utterance or the entire utterance can be played to help identify the phonetic events. Different visual displays of the speech signal can also be created in the layout. Although it is still necessary to have an experienced acoustic-phonetician to enter the phonetic events, the locations of the phonetic events need not be entered. The amount of man time saved is about one order of magnitude.

Figure 5-3 shows the results of the automatic alignment and how the acoustic landmarks can be adjusted. At the bottom of the Figure, there is an overlay of the spectrogram and the result of the alignment. The acoustic landmarks can easily be deleted, inserted, or adjusted by simply clicking the appropriate mouse buttons at the corresponding dashed line.



Speaker: Victor W. Zue
Sampling Rate: 16000.
Filename: Temporary-Utterance
Orthography: Automatic alignment of speech

Figure 5-1: A recording layout where utterances can be recorded and digitized

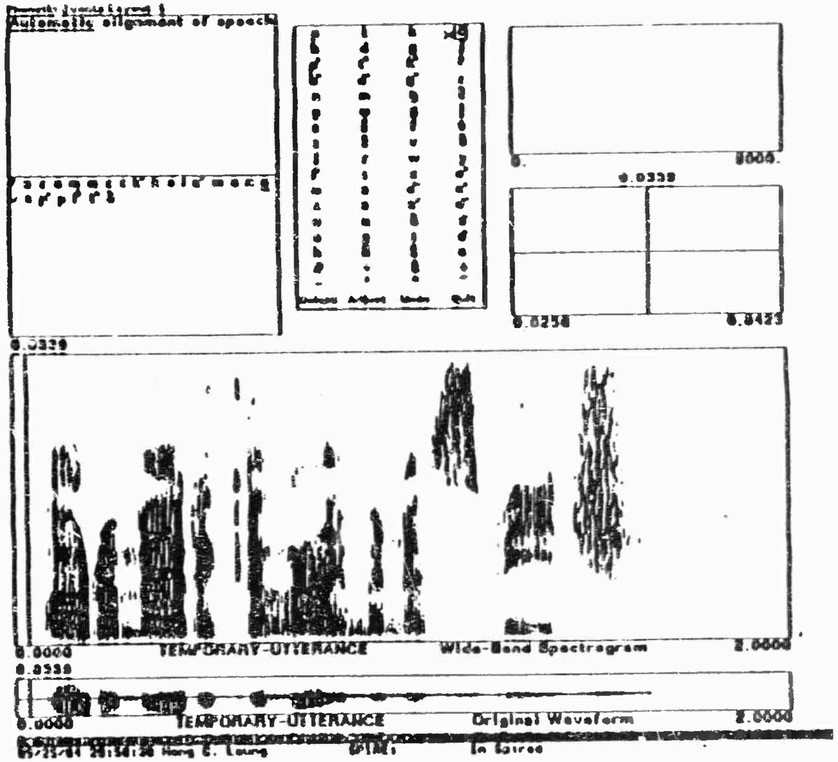


Figure 5-2: A phonetic event layout where phonetic transcriptions can be entered into the computer

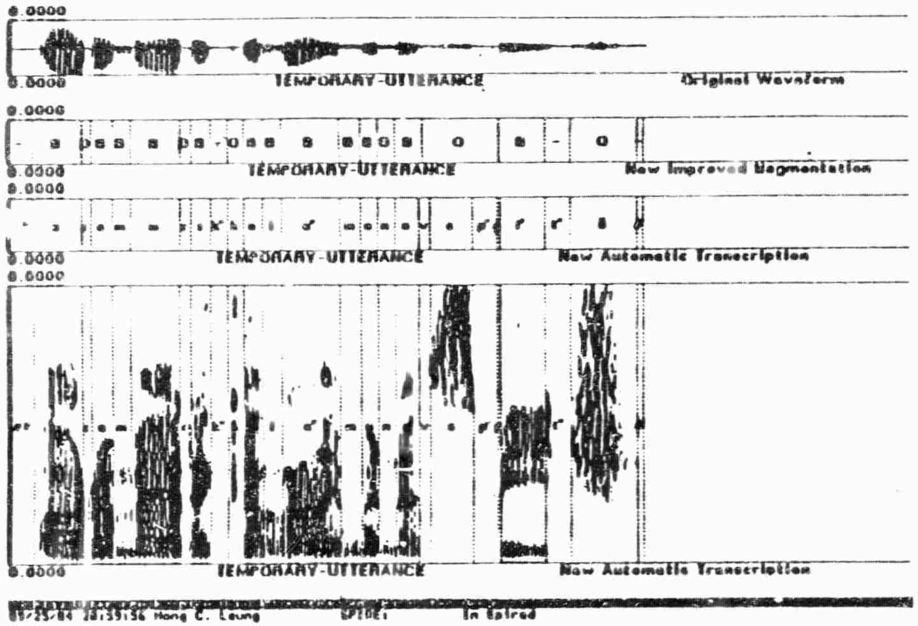


Figure 5-3: An alignment layout where the boundaries can be adjusted manually.

Appendix B

Evaluation Data

The evaluation data are randomly selected from the Harvard List of phonetically balanced sentences. There are forty distinct sentences, with a total of 100 tokens.

- The boy was there when the sun rose.
- A rod is used to catch pink salmon.
- The source of the huge river is the clear spring.
- Kick the ball straight and follow through.
- Help the woman get back to her feet.
- A pot of tea helps to pass the evening.
- Smoky fires lack flame and heat.
- The soft cushion broke the man's fall.
- The salt breeze came across from the sea.
- The girl at the booth sold fifty bonds.
- Hoist the load to your left shoulder.
- Take the winding path to reach the lake.
- Note closely the size of the gas tank.
- Wipe the grease off his dirty face.
- Mend the coat before you go out.

- The wrist was badly strained and hung limp

- The stray cat gave no birth to kittens.

- The young girl gave no clear response.

- The meal was cooked before the bell rang.

- What joy there is in living.

- The birch canoe slid on the smooth planks.

- Glue the sheet to the dark blue background.

- It's easy to tell the depth of a well.

- These days a chicken leg is a rare dish.

- Rice is often served in round bowls.

- The juice of lemons makes fine punch.

- The box was thrown beside the parked truck.

- The hogs were fed chopped corn and garbage.

- Four hours of steady work faced us.

- A large size in stockings is hard to sell.

- The slush lay deep along the street.

- A wisp of cloud hung in the blue air.

- A pound of sugar costs more than eggs.

- The fin was sharp and cut the clear water.

- The play seems dull and quite stupid.

- Bail the boat to stop it from sinking.

- The term ended in late June that year.

- A tusk is used to make costly gifts.

- Ten pins were set in order.

- The bill was paid every third week.

References

[Chamberlain 83]
Chamberlain R.M. and Bridle J.S.
ZIP: A Dynamic Programming Algorithm for Time-aligning Two Indefinitely Long Utterances.
ICASSP, 1983.

[Chen 84]
Chen, F.R. and Zue, V.W.
Application of Allophonic and Lexical Constraints in Continuous Digit Recognition.
IEEE International Conference on Speech Acoustics and Signal Processing, Paris, France, 1984.

[Cole 83]
Cole et al.
Performing Fine Phonetic Distinctions: Templates vs. Features.
Symposium On Invariance and Variability of Speech Processes, Massachusetts Institute of Technology, October, 1983.

[Hohne 83]
Hohne et al.
On Temporal Alignment of Sentences of Natural and Synthetic Speech.
IEEE Transactions on Acoustics, Speech, and Signal Processing Vol.ASSP-31 No.4, August, 1983.

[Huttenlocher 84]
Huttenlocher, D.P.
Acoustic-Phonetic and Lexical Constraints in Word Recognition: Lexical Access Using Partial Information.
S.M. Thesis, Massachusetts Institute of Technology, 1984.

[Itakura 75]
Itakura F.
Minimum Prediction Residual Applied to Speech Recognition.
IEEE Trans. Acoustic, Speech, and Signal Processing, Vol.ASSP-23, February 1975, 1975.

[Lamel 81]
Lamel L.F. et al.
An Improved Endpoint Detector for Isolated Word Recognition.
IEEE Trans. Acoustic, Speech, and Signal Processing. Vol ASSP 29, No 4, August 1981, 1981.

[Lennig 83]
Lennig M.
Automatic Alignment of Natural Speech with a Corresponding Transcription.
Speech Communication, 11th International Congress on Acoustics Toulouse, July, 1983.

[Lowry 78]
Lowry, M.R.
Automatic Labeling of Speech from the Phonetic Transcription.
S.M. Thesis, Massachusetts Institute of Technology, 1978.

[Miller 54]
Miller, G.A. and Nicely, P.E.
An Analysis of Perceptual Confusions Among Some English Consonants.
Journal of the Acoustical Society of America, Vol.27, No.2, 1954.

[Peterson 52]
Peterson, G.E. and Barney, H.L.
Control Methods Used in a Study of the Vowels.
Journal of the Acoustical Society of America, Vol.24, No.2, 1952.

[Rabiner and Schafer 78]
Rabiner L.R. and Schafer R.W.
Digital Processing of Speech Signals.
Prentice Hall, 1978.

[Rabiner 75]
Rabiner L.R. et al.
Applications of a Nonlinear Smoothing Algorithm to Speech Processing.
IEEE Trans. Acoustic Speech, and Signal Processing, Vol.ASSP-23, No.6, December 1975, 1975.

[Shipman 82]
Shipman, D.W. and Zue, V.W.
Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems.
IEEE International Conference on Speech Acoustics and Signal Processing, Paris, France, 1982.

[Tou and Gonzalez 74]
Tou J.T. and Gonzalez R.C.
Pattern Recognition Principles
Addison-Wesley, 1974.

[Wagner 81]
Wagner M.
Automatic Labelling of Continuous Speech with a Given Phonetic Transcription Using Dynamic Programming Algorithms.
ICASSP, 1981.

[Winston 84]
Winston P.H.
Artificial Intelligence.
Addison Wesley, 1984.

[Zue 80]
Zue, V.W. and Schwartz R.M.
Acoustic Processing and Phonetic Analysis.
Trends in Speech Recognition, Prentice Hall, 1980.

LEXICAL STRESS AND ITS APPLICATION IN LARGE VOCABULARY SPEECH RECOGNITION [1]

by
Ann Marie Aull
and
Victor W.Zue

Room 36-541
Department of Electrical Engineering and Computer Science
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

[1] Paper presented at the 108th Meeting of the Acoustical Society of America, Minneapolis, Minnesota, Oc'ober 10, 1984. The research was supported by the Office of Naval Research under Contract Number N00014-82-K-0727 and '.e System Development Foundation.

The study reported in this paper is concerned with the determination of lexical stress for isolated words from the acoustic signal. It is motivated by two observations. First, it has long been suggested that stressed syllables represent **islands of reliability** where the acoustic cues for phonetic segments are much more robust. Evidence for this observation has come from diverse sources. For example, phoneme-monitoring experiments for human speech perception have shown that reaction time is shorter for sounds in stressed syllables than for those in unstressed syllables. Analysis of human spectrogram reading results also indicates that accuracy is higher for sounds around stressed syllables. In addition, automatic speech recognition "front-ends" typically recognize phonemes around stressed syllables more accurately, again suggesting that the acoustic cues in these regions are more reliable. As an illustration, consider the spectrograms for the word pair CONtract/conTRACT shown in this figure. We can see that the characteristics for vowels and consonants around stressed syllables are more distinct.

A second reason for investigating lexical stress stems from the results of a set of experiments, suggesting that there are strong constraints on the allowable sound patterns in the English language. These studies have shown that, when words in a lexicon are represented in terms of broad manner classes, the number of words sharing the same broad-class pattern is often very small. In fact, when the broad class representation is augmented with stress pattern information, the number of word candidates for a given representation is further reduced. As an example, the two words "campus" and "compose" shown in this figure have the same broad class representation [STOP, VOWEL, NASAL, STOP, VOWEL, STRONG FRICATIVE], but they can be differentiated by the stress patterns.

Another finding of the studies we just cited is that, even if all the phonemes can be determined without error, stressed syllables still provide more lexical constraints than the unstressed ones. All these results seem to suggest that determination of the stress pattern of a word is potentially useful for speech recognition. We should emphasize at the onset that we are only interested in lexical stress, namely the stress pattern of words spoken in isolation. Sentential stress adds another level of complexity to the problem, and is not being addressed in our study.

Our present study focuses on two related issues. First, considering prosodic information as a separate source of knowledge, we investigate the amount of lexical constraint provided by stress information alone. In other words, we want to know by how much can the number of word candidates be reduced if only the stress patterns of words are known. Second, we implement a system that derives the stress information of a word, based on a set of measurements made from the acoustic signal.

In order to determine the lexical constraints provided by stress information, we created a lexicon from the Merriam-Webster Pocket Dictionary, consisting of all the two-, three-, and four-syllable words. The words in the lexicon were then mapped into their corresponding stress pattern classes. For this study, we allowed each word to have only one pronunciation and one stress pattern. We adopted a three-level convention, stressed, unstressed, and reduced, where a word must have one and only one stressed syllable. The results of the study are summarized in this figure. Looking at the left-hand column of numbers, we see that knowledge of the number of syllables of the words will give an expected class size equal to approximately 37% of the size of the lexicon. When the stress pattern is completely known, that is the correct number of syllables and the correct assignment of stress for each syllable, as shown in the middle column, the expected class size is reduced to 19%. In other words, we can expect to reduce the word candidates by a factor of five from stress information alone. Comparing the middle column to the right-hand one, it is interesting to note that knowledge of just the number of syllables and the location of the most-stressed syllable provides about as much constraints as the entire stress pattern.

Having determined the constraining power provided by stress information, we proceeded to develop an algorithm to automatically derive the stress pattern from the acoustic signal.. We note that the acoustic correlates of stress have been studied extensively by many researchers in the past, some of these studies are shown in this figure. Most of the studies have found that lexical stress is well correlated with the duration, the fundamental frequency value, and the intensity of the syllable nuclei. Lieberman in fact developed a system to automatically determine the stress pattern of bi-syllabic words. Prior to the actual development of the stress determination system, we also studied the acoustic correlates of stress based on a database of 350 words, spoken by 7 talkers, 3 male and 4 female. Our measurements generally agree with those found by previous researchers. There are two observations worth noting. First, we found that prepausal lengthening is an effect that must be properly compensated. Second, the measurements were nearly as effective in separating stressed syllables from unstressed ones if sonorants adjacent to the vowels were also included. This second observation is important, since it is sometimes difficult to delineate sonorants from adjacent vowels automatically.

The structure of the system that we have developed for stress determination is shown in the next figure. The input to the system is the acoustic signal, digitized at 16 kHz. The system has two main components. The first is a syllable detection component which establishes the sonorant regions of the syllables. Next, the stress algorithm examines the syllables within the word and derives a stress pattern. Thus, each derived syllable unit is labeled as stressed, unstressed, or reduced.

The next two figures discuss the system components in more detail. Syllable detection is accomplished in two stages. There is an initial segmentation which provides a description of the signal in terms of broad classes, such as sonorants, obstruents, etc. This classifier, developed by Leung and Zue, uses a number of speech parameters to classify the signal on a frame by frame basis. The classifier uses a number of sub-classifiers, arranged in a binary decision tree. At each node in the tree, a k-means clustering algorithm classifies each frame into one of two categories based on a specially selected feature set. Once sonorant regions are established, the second stage of the syllable detection further examines these regions for possible syllable boundaries, such as vowel/vowel or vowel/sonorant.

As an example of the syllable detection procedure,this next figure shows spectrograms and the initial segmentation boundaries marked in blue for the words "Massachusetts", "yellow", and "create". For the first word, "Massachusetts", each sonorant region derived by the initial classifier corresponds to a syllable unit. No further work is necessary. However, "yellow" and "create" have an intervocalic sonorant and a vowel/vowel transition, respectively, not detected by the initial sonorant classification. The second stage of syllable detection examines the contextual information within these sonorant regions and establishes additional syllable boundaries shown in red.

The second component determines the stress pattern from a set of measurements made from each syllable. Measurements include duration, energy in different frequency bands, and fundamental frequency, similar to previous work. In addition, a spectral change measure reflects the amount of spectral stability over time. Duration is measured by including the vowel and any surrounding sonorants as determined by the syllable detection component. An average of energy and spectral change is computed over each syllable. Finally, the maximum of fundamental frequency over each syllable is used.

The next phase of stress determination is the decision algorithm. First, the above parameters are compensated, such as for prepausal lengthening, and normalized. Each syllable is then associated with a feature vector. The assignment of stress is based on a relative comparison of the feature vectors for all the syllables in a given word, and does not rely on an absolute targets for stressed and unstressed syllables obtained through a training set. As a result, the algorithm provides an implicit timing normalization for differences in word lengths and number of syllables, and is relatively insensitive to inter- and intra-speaker variabilities. The "optimum" parameter value across the word for each parameter forms an extremum in the feature space. A Euclidean distance from each syllable vector to the extremum is computed. The syllable with the minimum distance is then labeled as stressed. Reduced decisions are made by reexamining the duration and energy in the remaining unstressed syllables. Thus, the output of the system is a time-aligned stress pattern for the input word.

The system is evaluated on a corpus of 1600 isolated words, consisting of 2,3,4, and 5 syllable words. There are a total of 4500 syllable tokens in the corpus. In addition, 11 speakers, both male and female, are included in the evaluation.

The results of the evaluation can be divided into three criteria, in increasing order of difficulty. The first is the determination of the stressed syllable, even in the presence of syllable detection errors. In this case, 2% of the words do not have the stressed syllable labeled correctly. The next criterion imposes the additional constraint that the number of syllables must also be correctly identified. With these more stringent requirements, the error rate increases to 10% of the corpus. Finally, the stipulation that the stress pattern must be correct demands correct syllable and stress identification, as well as no confusion between unstressed and reduced segments. These additional constraints increases the error rate to 13% of the corpus. Comparing these results, we see that most of the errors made by the system can be attributed to inaccurate syllable detection, rather than stress assignment. Recall earlier, we have shown that the lexical constraints provided by knowing the entire stress pattern is almost the same as knowing the number of syllables and the most stressed one. Thus, the performance deterioration from 10% to 13% may not be too serious.

In summary, we have determined that lexical stress information can provide strong constraints towards word candidate reduction. The algorithms that we have developed can automatically and reliably determine the stressed syllables in isolated words. By identifying the stressed syllables, the system can provide pointers to regions where the acoustic information is presumably robust, and thus improve the performance of phonetic recognition. A majority of the errors for stress pattern determination can be attributed to an ambiguity associated with the number of syllables for certain polysyllabic words, a task often difficult for humans as well. By incorporating syllable reduction and deletion rules and thus increasing the number of alternate pronunciations, we believe this system can eventually be incorporated in a large vocabulary speech recognition system.

Figures used in talk follow this page.

Motivation I

Acoustic characteristics of speech sounds are more robust around stressed syllables.

Evidence from:

- Speech Perception (e.g. Cutler and Foss, 1977)
- Spectrogram Reading (e.g.. Klatt and Stevens, 1973)
- Speech Recognition Systems (e.g. Lea, 1980)

Motivation II

Stress information provides strong constraints for lexical access (Zue and Huttenlocher, 1983).

[Stop Vowel Nasal Stop Vowel Strong-Fricative]

Stressed syllables provide more lexical constraints than unstressed ones.

How Much Lexical Constraint Does Stress Information Provide?

Corpus: Merriam-Webster Pocket Dictionary

Size: approximately 15,000 (all two- through four-syllable words)

	# of Syl.	# of Syl. & Stress Pattern	# of Syl. & Location of Stressed Syl.
Expected Class Size	5590 (37%)	2915 (19%)	3180 (21%)

Past Studies on Lexicai Stress

Acoustic Correlates of Stress:

(e.g. Fry, 1955; Bolinger, 1958; Mcrton & Jassem, 1965; Lehiste, 1970)

- Duration
- Fundamental frequency
- Intensity

Recognition System: (Lieberman, 1960)

System Overview

speech
signal

Syllable
Detection

Stress
Determination

stress
pattern

Syllable Detection

- Initial Classification to Establish Sonorant Regions

  *Classifier identifies broad phonetic categories (Leung & Zue, 1984)

  *Frame-by-frame classification from speech parameters

  *Binary tree-like structure

  *K-means clustering algorithm

- Further Investigation of Sonorant Regions for Possible Vowel/Vowel and Vowel/Sonorant Transitions

Stress Determination

- Parameters:

  *duration

  *energy in different frequency bands

  *fundamental frequency

  *spectral change

- Algorithm:

  *Measurements are normalized and compensated

  *Comparisons are made among syllables within a word

  *Euclidean distance measured from extrema in the feature space

Evaluation

Corpus

# of Words	1600
# of Syllables	4500
# of Speakers	11 (6M, 5F)

Evaluation

Results

Criterion	Error Rate (%)
Stressed Syllables	2
# of Syllables & Location of Stressed Syllable	10
Stress Pattern	13

An Acoustic Study of Nasal Consonants in American English

James R. Glass and Victor W. Zue

36-541
Department of Electrical Engineering and Computer Science
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139

This paper was presented at the 108th Meeting of the Acoustical Society of America in Minneapolis, Minnesota, October 9, 1984. The research was supported by the Natural Sciences and Engineering Council of Canada and by the Office of Naval Research under contract N00014-82-K-0727.

I'd like to talk to you today about some work we have been doing which is concerned with the acoustic characteristics of nasal consonants in American English. There are actually two goals that we are trying to achieve in this work. First, we are attempting to quantify some of the acoustic characteristics of the nasal consonants in American English. Although nasal consonants have been studied extensively in the past by many researchers, we feel that we can still contribute to this body of knowledge by trying to quantify some of the basic acoustic properties of nasal consonants from a reasonably sized database. The second goal of this work is to apply any robust acoustic characteristics which we are able to establish to the field of automatic speech recognition.

The acoustic analysis of the nasal consonants was made from a database specifically created for this work. The database was based on a carefully constructed corpus of some two hundred words which was designed to contain nasal consonants in many different contexts. Thus the corpus contained nasal consonants in prevocalic, medial, and post-vocalic contexts. It contained nasal consonants in both singleton environments and in consonant clusters. The corpus also contained minimal pairs of words which could be used to study the subtle differences between nasalized and non nasalized vowels, or between nasal consonants and other sounds which might be confused with a nasal consonant. The corpus also contained minimal pair words which could be used to study the differences between nasal consonants with different place of articulation.

Once the corpus had been designed the database was created. Three male and three female speakers each read the words of the corpus which were embedded in a carrier phrase. All utterances were digitized at 16 kHz and the phonetic transcriptions were manually time aligned with the waveforms. The final database contained over 1200 words.

The main point to make about the acoustic analysis of the database is that apart from the original time alignment of the phonetic transcription, all measurements and analyses were done automatically by machine. One advantage to this procedure is that any acoustic characteristic that can be shown to be robust can be immediately used in an automatic speech recognition system. Another advantage of automatic analysis is that it allows the analysis of a large amount of data in a reasonable amount of time. For instance, the size of the database was limited primarily by the time it took to complete the time alignment. Automatic analysis must be done carefully however, or one runs the risk of adding measurement noise into the distributions of the data.

In our analysis of the nasal consonants we were interested in three major areas: the nasal murmur or the period of oral closure, the period of nasalization in an adjacent vowel, and the period of transition between the nasal murmur and an adjacent vowel. Work has been completed on the first two areas and we are presently studying the transition region. Due to the time constraint, I will limit my discussion in this talk to some of the basic findings of our work with the nasal murmur. In addition, I will discuss some preliminary results we have obtained with automatic nasal consonant detection using acoustic information about the nasal murmur only.

The first thing that we studied about the nasal murmur was duration. We found, as did many previous studies of nasal consonants, that the duration of the nasal murmur is strongly influenced by phonetic context. Figure 1 contains a statistical summary of the duration of nasal murmurs in either a singleton environment, or in a cluster with another consonant. For each group, the mean is indicated by a filled circle, and the vertical bars represent one standard deviation. The open circles indicate the minimum and maximum values for all of the samples. As shown in these statistics, consonant clusters tend to reduce the duration of the nasal murmur. However when we look a little closer we find that the actual effect depends on the voicing characteristics of the adjacent consonant. Thus, the nasal murmur is shortened when it is in a cluster with a voiceless consonant, and is lengthened when it is a cluster with a voiced consonant. This result was found to be true for both word-initial clusters, as in *smack*, and word-final clusters, as in *can't*.

Although these results are fairly robust, their immediate application to speech recognition is limited since one would need to know the exact context to be able to apply this information.

The second acoustic characteristic which was quantified was the energy of the nasal murmur. The measurement procedure is illustrated in figure 2. Above the spectrogram of the word *hammock* are the zero-crossing rate, the total energy, and the low-frequency energy. The energy difference between the nasal and the adjacent vowel was calculated by subtracting the average total energy in the nasal murmur from the average total energy in the adjacent sonorant. In the bottom part of the figure you can find a histogram of this energy difference, plotted in dB. Since this energy difference is almost always positive, we conclude that the nasal murmur is consistently weaker than an adjacent vowel by an average of 10 dB. As a comparison, we also show the same measurement for liquids and glides. Although the distributions overlap somewhat,it can be seen that liquids and glides tend to have less of an energy difference than nasals, i.e., they have more relative energy than do nasal consonants. Thus, from a speech recognition point of view, this measurement may help to distinguish nasals from liquids and glides. In fact, this is one of the measurements that we used in a recognition experiment that we will describe shortly.

The next acoustic characteristic which was examined was the spectral characteristics of the nasal murmur. For analysis purposes we calculated statistics based on a cepstrally smoothed spectra created from the preemphasized waveform. The spectra were all normalized with respect to total energy so that we did not have to be concerned with an offset. During analysis we also restricted ourselves to analysing the spectra of one speaker at a time since we found that the spectra, primarily at frequencies above 1000 Hz, were highly speaker dependent. This may not be surprising, since the size of the nasal and sinus cavities can vary greatly from speaker to speaker. Statistics were gathered by collecting multiple spectra from all of the nasal murmurs. The top of figure 3 shows multiple spectra for *m* for one speaker after energy normalization. We see that there are some common characteristics among the many spectra, and we tried to capture the essence by averaging the spectra. We found little difference between the average spectra obtained from multiple spectra and that obtained from an average spectra for each murmur. The two bottom displays illustrate average spectra for an intervocalic *m* for one speaker using these two different techniques. The thick line is the mean spectral shape and the outer two lines are one standard deviation away. As can be seen, the average spectral shape is very similar. The standard deviation of the multiple spectra, shown on the left, is larger which is to be expected since more spectra were included. The fact that the two averaging technique yield similar results points out the fact that the spectral characteristics of the nasal murmur are quite stable against time, especially at low frequencies. For the remainder of this talk, we have used the average spectra obtained from the multiple spectra technique.

Figure 4 shows an average spectra for an *n* for one speaker. On top of this we have drawn the average spectra of *m* for the same speaker and also the average spectra of an *ng*. In general we found that the spectral shapes of the nasal murmur were highly speaker dependent. Further, although subtle differences could be detected between the three nasal consonants for any given speaker, all three nasal consonants tended to have similar spectral shapes as we can see here. This observation is in agreement with that made by Fujimura, who also find little differences among the spectra of the three nasal murmurs. Finally we found that the spectral shape of the nasal murmur was relatively unaffected by phonetic context.

In general we found that the nasal murmur spectra were characterized by a low frequency energy which dominated the spectrum. As illustrated in figure 5, this low frequency energy was nearly always centered between 200 and 350 Hz. This characteristic is very strong in that if a spectra does not have a resonance centered in this frequency range then it is likely not from a nasal murmur.

As previously mentioned, this low energy dominates the overall spectrum of the nasal murmur. We found that the normalized resonance amplitude ranged from 20 to 30 dB for most spectra. Another characteristic of the nasal murmur was a fairly abrupt decrease of energy in the frequencies immediately following the low resonance. This drop which we have labeled as the resonance height was on average about 10 dB.

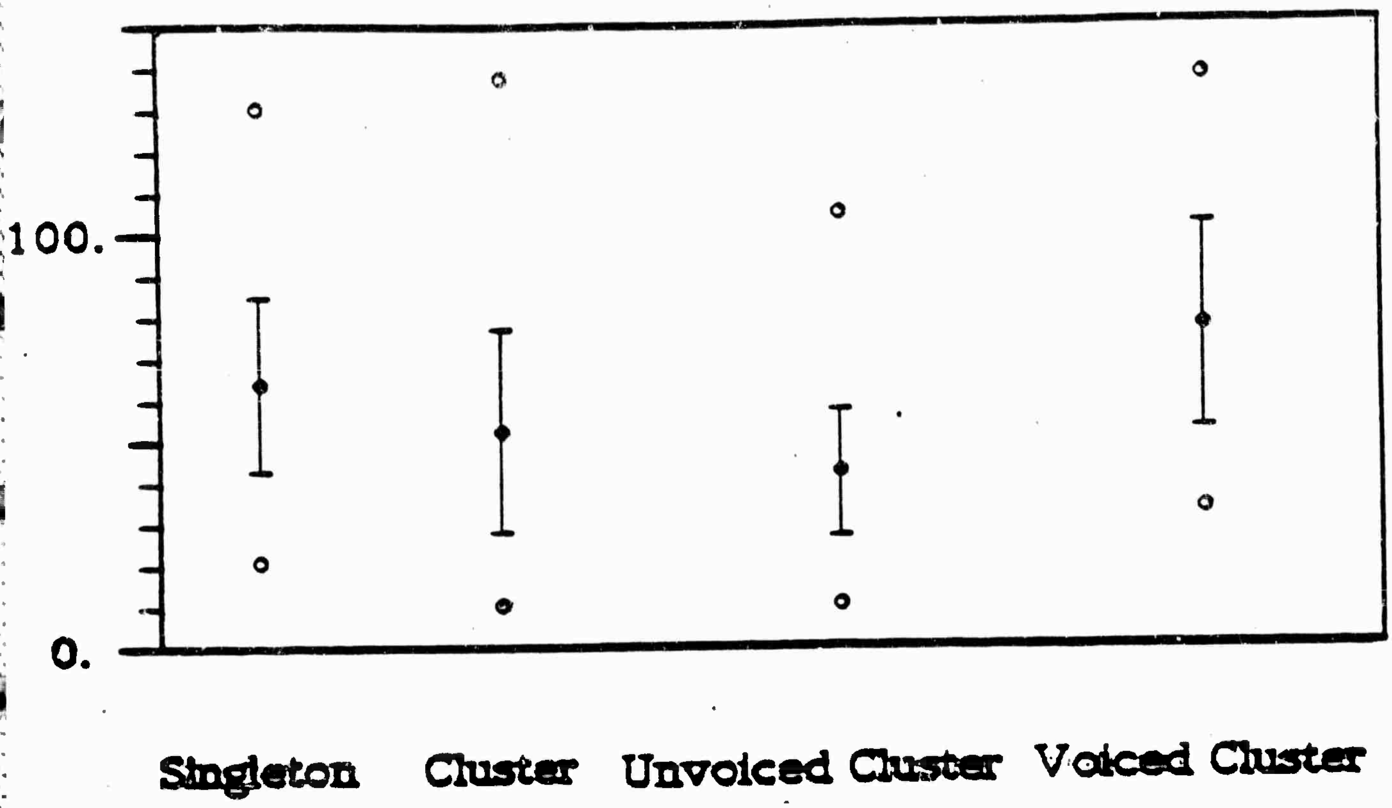
After observing some of the basic characteristics of the nasal murmur, we were interested in determining how well we could discriminate between nasal consonants and impostors such as liquids, glides, voice bars and voiced fricatives. For the time being we have restricted ourselves to using information in the nasal murmur alone. Further, we wished to utilize only measurements that are common across different speakers.

Our strategy was to combine five simple measures into a maximum likelihood decision making process. The five measures included the energy difference, the percentage of the time that there was a low frequency resonance centered between 200 and 350 Hz in the nasal murmur, the average amplitude of this resonance in the murmur, the average height of the resonance all of which have been mentioned previously. We also included a simple measure of the steadiness of the spectrum which was based on normalized low frequency energy.

As a first step at evaluation, the recognition task was performed using the utterances of the database for all six speakers. There were 520 nasal murmurs and 695 impostor sounds. In the recognition of one murmur, all other murmurs were utilized for training. The results indicate that the system can detect the presence of a nasal consonant 94% of the time and can detect an impostor 81% of the time. [1] The next step will be to evaluate the system on a different database containing a large number of different speakers.

In summary, we have found some distinct characteristics of nasal murmurs and we are encouraged by the preliminary results of the recognition task. We feel that our study supports the notion that a better understanding of the acoustic properties of speech sounds will help us to perform better speech recognition.

[1] The authors ask to be consulted before these results are quoted since they were obtained at such a preliminary stage in the evaluation.



Nasal Murmur Duration

- Average Duration by Environment:
  1. Singleton: 65 msec
  2. Cluster: 55 msec
  3. Voiceless Cluster: 40 msec
  4. Voiced Cluster: 75 msec

Figure 1

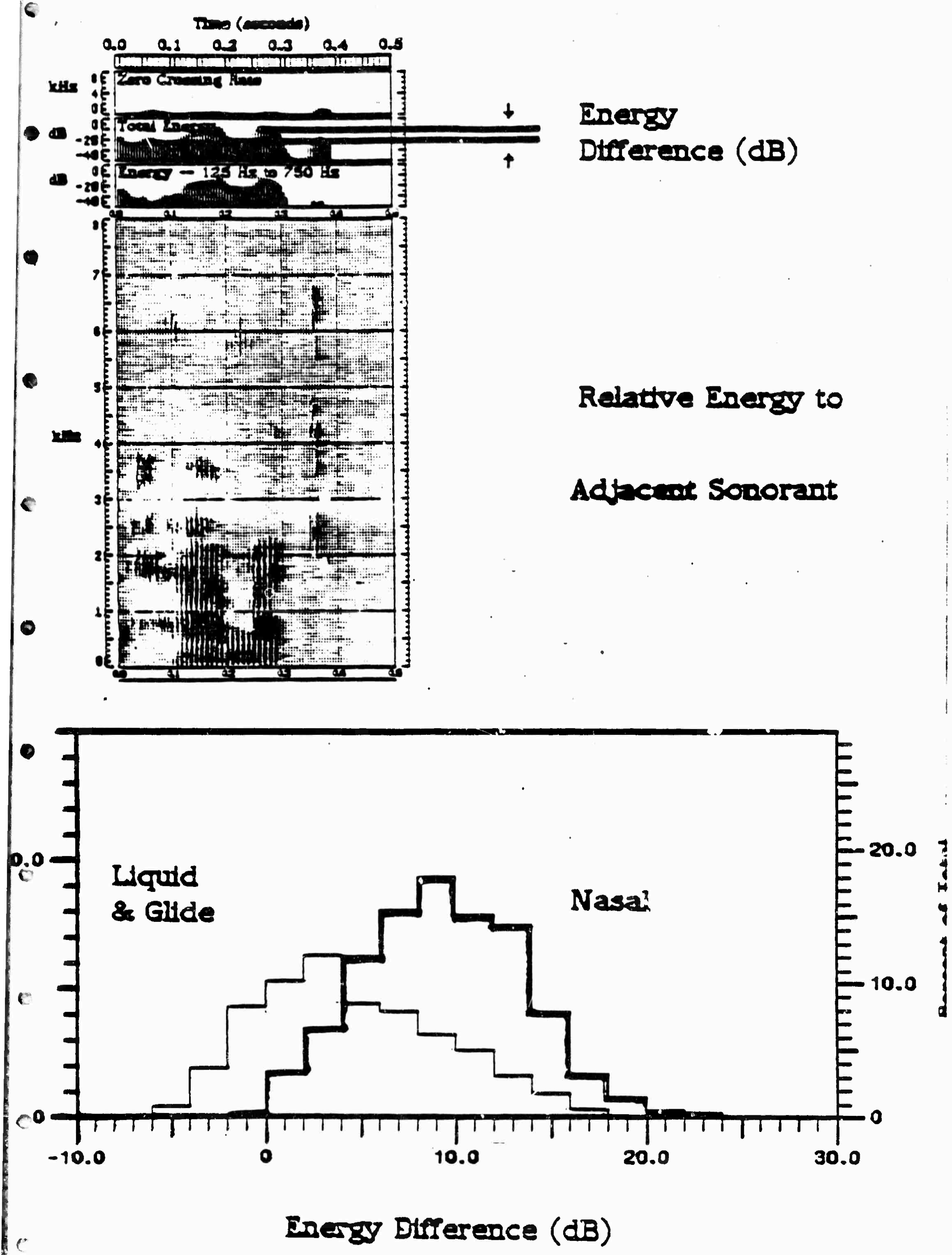
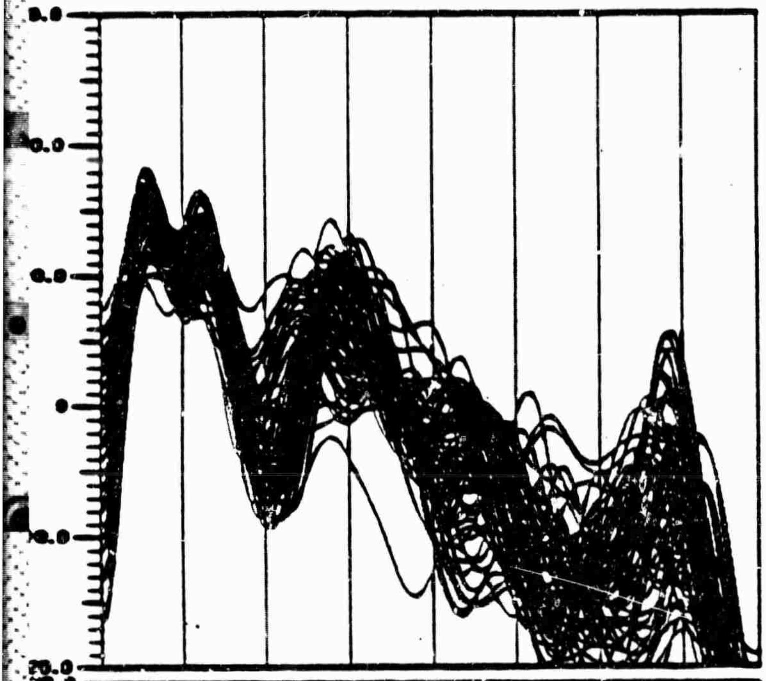
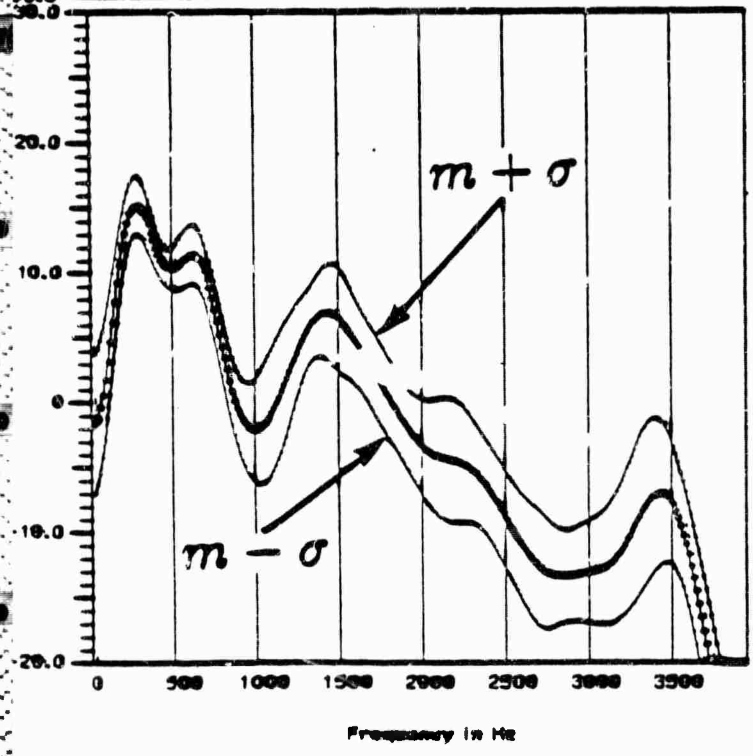

Time (seconds)
0.0
0.1
0.2
0.3
0.4
0.5
kHz
Zero Crossing Rate
dB
Total Energy
dB
Energy -- 125 Hz to 750 Hz
kHz
Energy
Difference (dB)
Relative Energy to
Adjacent Sonorant
Liquid
& Glide
Nasal
0.0
0
-10.0
0
10.0
20.0
30.0
20.0
10.0
0
Percent of Total
Energy Difference (dB)

Figure 2

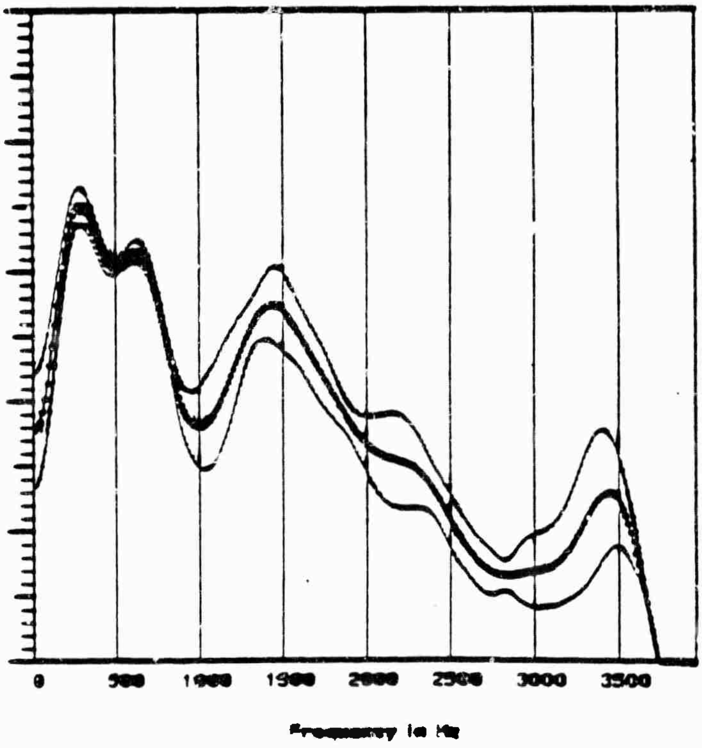
Spectral Statistics



- pre-emphasized waveform
- smoothed spectrum
- energy normalized
- speaker dependent
- multiple tokens from murmurs

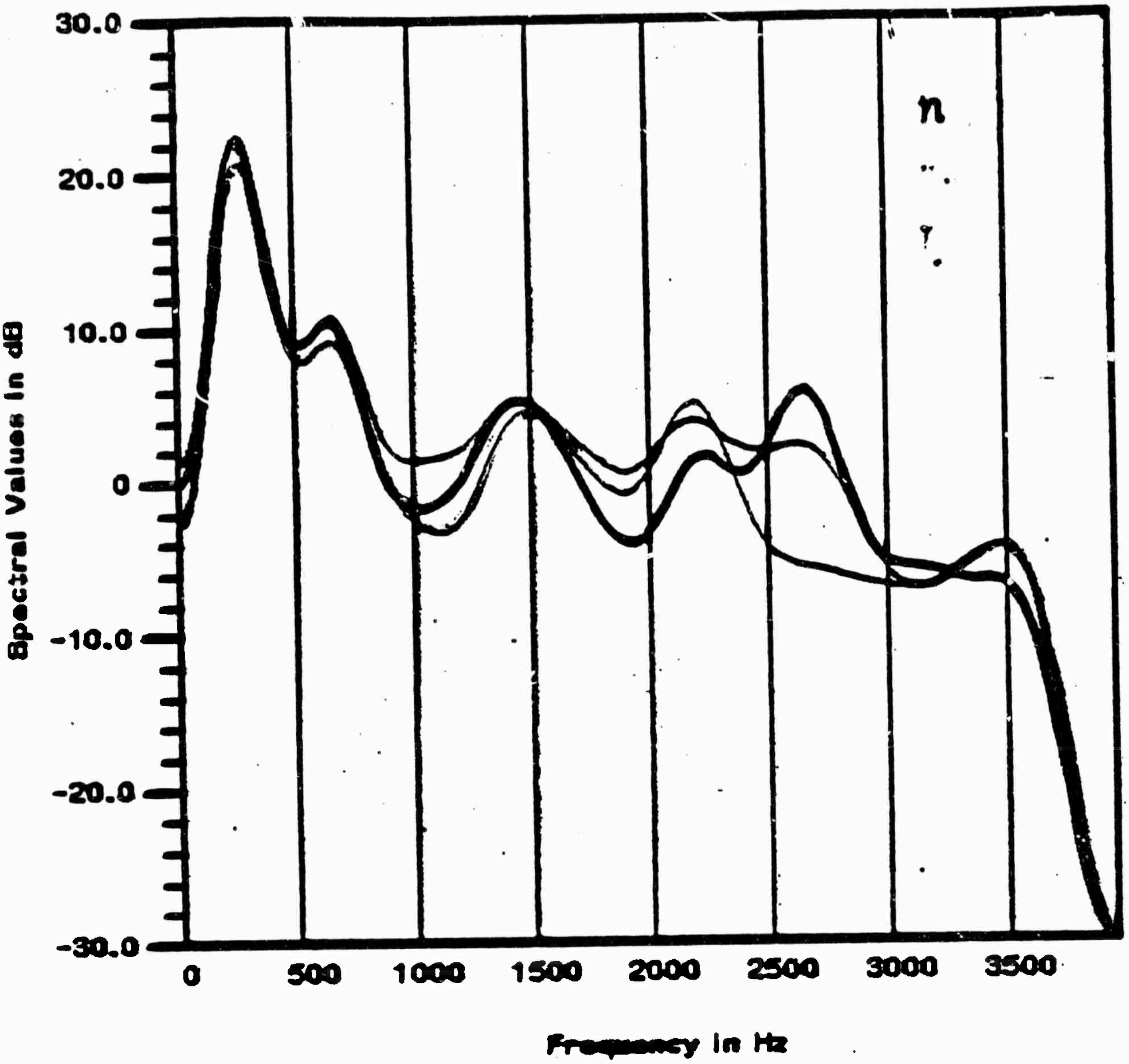


Multiple Spectra



Average Spectrum

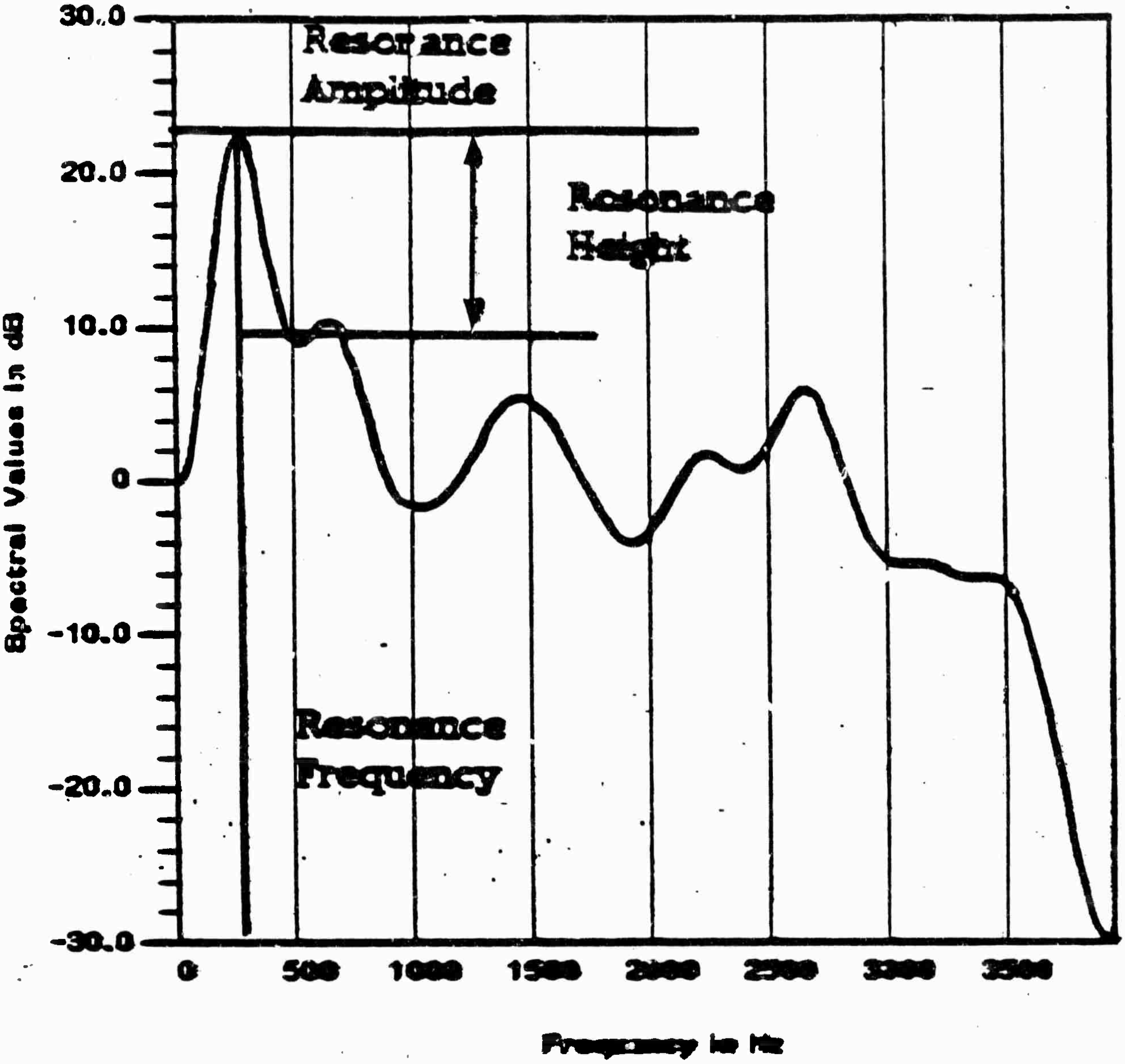
Figure 3



Spectral Characteristics

- spectral shape speaker dependent
- similar shapes for all three consonants
- little effect due to context

Figure 4



Spectral Characteristics

- energy centered between 200 - 350 Hz
- normalized amplitude average: 25 dB
- resonance height average: 10 dB

Figure 5

Paper presented at ICASSP 85, Tampa, Florida, March 28, 1985.

DETECTION OF NASALIZED VOWELS IN AMERICAN ENGLISH*

James R. Glass and Victor W. Zue

Department of Electrical Engineering and Computer Science
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

This study is concerned with the quantification of acoustic measures that characterize nasalized vowels. It is motivated by the fact that the detection of nasalized vowels is useful in speech recognition, since regions can be identified in the speech signal where a nasal consonant may be present, and where the vocal tract resonances are distorted. Our study consists of several steps. First, an acoustic study was performed using utterances from a large database in order to propose potential measures of nasality. Next, automatic algorithms were developed to extract these measures, and their utilities were established through examination of a large amount of data. Finally, recognition experiments were performed using these measures. The system detected nasalized vowels with an accuracy of approximately 74%, when tested on one speaker at a time, and trained on the speech of the remaining speakers in the database.

INTRODUCTION

This study is concerned with the acoustic analysis of nasalized vowels and the subsequent development of algorithms for their recognition. In American English, vowels adjacent to nasal consonants are often nasalized. The presence of a side-branch for sound transmission introduces additional poles and zeros into the vocal tract transfer function. As a consequence, the short-time spectra of nasalized vowels often exhibit extra *nasal* formants, or a broadening of the vowel formants, typically in the first formant region.

There are several reasons why the detection of nasalization in vowels would be useful for automatic speech recognition. First, vowel nasalization provides important information regarding the presence of a nasal consonant, especially in contexts where the nasal murmur has been shortened considerably or is absent altogether, as in words like *smack*, or *can't*. In these cases, the most reliable cue for the presence of a nasal consonant is often in the degree of nasalization of adjacent vowels. Second, nasalized vowels present a problem for formant tracking algorithms, since correct assignment of the spectral peaks to the vocal tract formant frequencies becomes more difficult due to the pole-zero interactions. If the regions of nasalization were known, different tracking strategies could be employed.

The purpose of our study is two-fold. First, we would like to establish the acoustic properties that distinguish nasalized vowels from oral vowels. We accomplish this by drawing from knowledge gained from past studies to propose, and evaluate measures of nasality using a specially designed database. Second, the measurements found to be reliable are incorporated into a recognition algorithm for speaker-independent nasalized vowel detection.

There is a vast amount of literature spanning the last twenty-five years dealing with the analysis, synthesis, and perception of nasalized vowels. Extensive acoustic studies have shown that nasalized vowels are characterized by the presence of a low-frequency formant and antiformant, as well as additional weaker spectral peaks in the spectral valleys between oral formants [1, 4]. The low-frequency pole/zero pair often broadens or splits the spectral peak associated with the first formant. Synthesis experiments based on these findings have shown that the perceptually salient cue for nasalization appears to be the reduction of the relative prominence of the first formant peak [5]. These results, while important in leading to a better understanding of the acoustic characteristics of nasalized vowels, are often not directly applicable to speech recognition. Some of the data has not been presented in sufficiently quantitative form, and the measurements often rely on human intervention and interpretation. In many cases, the data has been gathered from restricted environments such as stressed consonant-vowel-consonant (CVC) syllables.

DATABASE DESCRIPTION

Our study utilizes a database originally collected for a separate study of the properties of nasal consonants in American English [2]. The corpus, containing over two hundred words, is designed to include nasal consonants in many different contexts. It contains nasal consonants, both in singletons and in clusters, that appear in syllable initial, medial, and final positions. Many of the words also form minimal pairs such as *cap/camp*, and *sack/snack*. Words in the corpus were embedded in a carrier phrase and recorded by six speakers, three male and three female, resulting in over 1,200 word tokens. All of the recorded words were digitized at 16 kHz, excised from the carrier phrase, and their phonetic transcriptions aligned with the speech waveform using the Spire system [7].

ANALYSIS ISSUES

An acoustic study of nasalized vowels in American English is complicated by several issues. First, speakers are free to nasalize a vowel in any phonetic context, since nasalized vowels are not distinguished phonemically from oral vowels in American English. As an illustration, consider the spectrograms shown in Figure 1. The left and middle panels contain the words *mack*, and *back* respectively, spoken by speaker A, whereas the right panel contains the word *mack*, spoken by speaker B. Careful listening of the vowel indicates that speaker A has nasalized the /ae/ in the word *back*. This is evident in the spectrogram by the presence of a low-frequency spectral peak below the first formant. In fact, many speakers frequently nasalize their vowels, irrespective of context. To be of use for nasal consonant detection, the acoustic study must establish measures which can automatically differentiate between vowels in a nasal consonant context and those not in a nasal consonant context.

* Research supported by the Natural Sciences and Engineering Research Council of Canada and by the Office of Naval Research under Contract N00014-82-K-0727

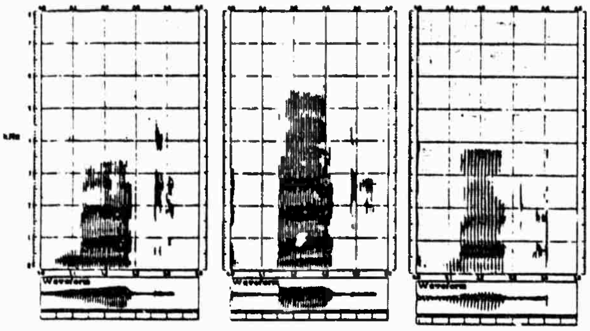


Figure 1: Spectrograms of the words *nack*, *hack*, and *mack*

Examples such as those shown in Figure 1 suggest that it is not possible to distinguish nasalized and oral vowels by phonetic context alone. In order to accurately determine the presence of nasalization, the acoustic signal must be augmented with other physiological measurement indicating the degree of nasal coupling. Without such additional measurements, one can only *infer* vowel nasalization by perceptual experiments, or by arbitrarily determined criteria. In this study, nasalized vowels are defined as *those vowels adjacent to a nasal consonant*, while non-nasalized vowels are *those vowels that are not adjacent to a nasal consonant*. Thus, our acoustic study can be viewed as an analysis of *relative* nasalization. The underlying assumption is that when vowels are next to a nasal consonant, they are nasalized more than they would be otherwise. This definition is directed more for nasal consonant detection than for general nasalization detection since there are clearly some vowels which are not in a nasal consonant context which are nasalized. However the results of the acoustic study are applicable to both areas.

Our definition for nasalized vowels is admittedly inadequate, since according to such a definition, vowels which are separated from a nasal consonant by an intervening consonant (as in *film*) will be classified as a non-nasalized vowel, whereas in all probability these vowels will be nasalized. Since the nature of these vowels is somewhat ambiguous, we have chosen to eliminate such tokens from our database in order to reduce the amount of noise they might cause in the measurement distributions. This excluded about 200 vowels from the acoustic analysis.

A second difficulty with a study of nasalized vowels is that different speakers nasalize to various degrees. Thus, one person's nasalized vowel could have the same characteristics as another's non-nasalized vowel. Again, referring to Figure 1, we see that the strong cues for nasalization in speaker A are barely present for speaker B. This phenomenon smears measurement distributions, and compounds the difficulties associated with speaker-independent nasalized vowel detection.

In our acoustic analysis, several procedures were adopted to control the influence of interspeaker variability. The analysis is conducted on a speaker-by-speaker basis in order to eliminate the speaker-dependent nature of vowel nasalization. In fact, the initial portion of the analysis is restricted to observing relative differences between minimal word pairs such as *skip/skimp* for each speaker, thus maximizing the opportunity for us to observe acoustic contrasts between nasalized and non-nasalized vowels.

A third factor complicating the acoustic analysis is the inherent dynamic quality of nasalized vowel spectra. Sometimes the spectral change is due to the time course and the varying degree of nasal coupling throughout the vowel production. Other times the dynamic characteristics can be attributed to the movements of other articulators, such as during the production of dipthongs. In either of these cases, the net effect is that the acoustic characteristics change with time.

While it may be possible to track significant characteristics of the vowel (such as the resonance frequencies) as a function of time, this method was not used because it was felt that such systems may be fragile, especially in nasalized vowels. Instead, each vowel was divided into *subsegments*, so that averaging procedures could be used in each subsegment to reduce measurement noise. At the same time however, changes between the different subsegments of the vowel, caused by increasing nasalization, would still be measurable. After some experimentation it was decided to use three subsegments in each vowel. Thus, whenever a measurement of some parameter was made on a vowel, there were three values returned. Each value represented an average of the parameter in one of the three, equally spaced, vowel subsegments. An added benefit of such a procedure is that, by comparing the measurements in the initial and final portions, one may be able to determine whether the vowel is preceded or followed by a nasal consonant.

ACOUSTIC STUDY

In the acoustic analysis, the goal was to establish differences between nasalized and non-nasallized vowels by comparing some form of average spectra of selected tokens from the database. On the basis of these observations, general discriminating properties could be proposed and quantified using all utterances of the database. The following sections describe the steps followed for the spectral analysis. A more detailed description of the acoustic analysis may be found in [3].

Spectral Averaging

For analysis, the speech signal was pre-emphasized, and spectra were computed from a windowed cepstrum [6]. The spectra were all normalized with respect to total energy so that individual energy offsets were eliminated. Statistics were gathered by averaging multiple spectra for nasalized and non-nasalized vowels of each speaker. Figure 2 shows average spectra for an /ae/ for a male speaker. The left panel presents a statistical summary of the normalized, smoothed spectra of the non-nasalized /ae/ of a male speaker. The right panel presents the corresponding nasalized /ae/ of the same speaker. The average spectral shape, shown by the dark line, is surrounded by lines which represent one standard deviation from the mean.

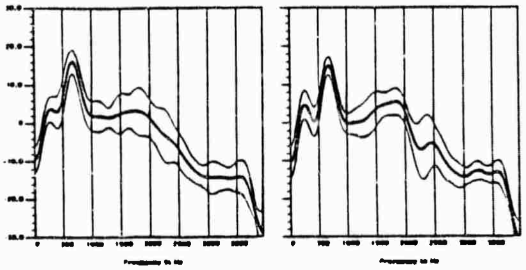


Figure 2: Average Spectral Shape of /ae/

Although the averaging procedure can potentially smear useful information, we nevertheless found these displays very informative. By comparing the vowel spectra such as those in Figure 2, we were able to establish general characteristics of nasalized vowels. For example, we found that the most noticeable difference between the nasalized and non-nasalized vowels was in the low frequency regions of the magnitude spectrum. Typically, non-nasalized vowels had one resonance in the first formant region, while nasalized vowels had two. However, many non-nasalized vowels also had an extra resonance in the low frequency region.[1] Thus, it is not always possible to distinguish nasalized from non-nasalized vowels by simply measuring how often there is an extra resonance in the first formant region of the vowel.

Another characteristic of nasalized vowels is that the extra resonance is noticeably more *distinct*, and is manifested in the spectrum in at least two ways. First, the magnitude of the extra resonance may increase relative to the first resonance of the vowel. This can be caused by the first resonance decreasing in amplitude, or the extra resonance increasing, or both. Second, the valley between the extra resonance and the first resonance may deepen. Thus, even if a non-nasalized and a nasalized vowel both happened to have an extra resonance, it may still be possible to discriminate between them by measuring the strength of the extra resonance relative to the first formant.

Another observed characteristic of nasality was a smearing of the first resonance itself. In fact, when an extra resonance was not present, as occasionally observed in a nasalized vowel, a measure of the spread of energy about the first resonance was found to be the best indication of nasalization.

By observing the characteristics of these averaged spectra on a subset of the database, we were able to propose measures that may signify the acoustic contrasts between the two classes of vowels. Due to the variability of the environment, none of the observed acoustic characteristics was present in a nasalized vowel at all times. However, taken in combination, we felt that these properties could help discriminate nasalized vowels from non-nasalized ones.

The next step of our acoustic study was to formalize our observations into a set of specific algorithms for automatic feature extraction, and to validate and quantify these measures by examining its statistical behavior on the entire database. Figure 3 illustrates a typical result of the analysis which compares a measure of the spectral spread of nasalized vowels versus non-nasalized vowels. The spread was calculated by computing the standard deviation of a spectral first moment, computed below 1000 Hz [3]. In the display, the dark lines are the distributions of nasalized vowels (695 samples), while the dashed lines are the distributions of non-nasalized vowels (500 samples). All values are in Hz. We can see that for this measure, the two classes of vowels have different, although overlapping distributions.

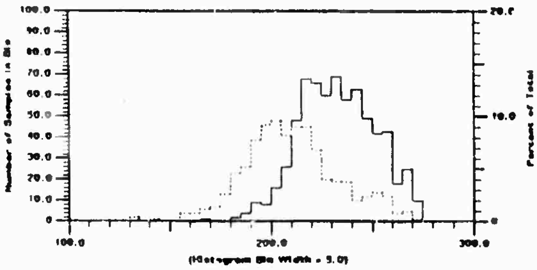


Figure 3: Histogram . Standard Dev'ation

RECOGNITION EXPERIMENT

After establishing some of the important acoustic characteristics of nasalized vowels, preliminary investigations were conducted to evaluate the potential use of these properties in speech recognition. Admittedly, these experiments do not realistically reflect the utility of the measurements, since the evaluations were made on the same database as the acoustic study. However, they do provide an indication of their potential for use in speaker-independent, speech recognition systems.

The Task

We have structured the recognition experiments as a set of *discrimination* tests. Thus in a typical experiment, the nasalized vowel detection system is given a test token and training data. The system must then classify the token as either next to a nasal consonant (nasalized), or not next to a nasal consonant (non-nasalized). Throughout these experiments, it is assumed that the boundaries of the vowels are known, although no knowledge of the presence or absence of a nasal murmur is used.

The Strategy

Our acoustic study produced several parameters, each potentially useful in characterizing a certain aspect of nasalized vowels. We have chosen to incorporate all these measurements into detection systems for the task in hand. Thus a given test token is associated with a set of *n* values, corresponding to a set of acoustic measurements made on the test token. If we consider the set of values as a vector in an n-dimensional space, we are faced with a multidimensional decision-making problem.

Although there are several possible decision making procedures available, a sum of individual log likelihoods was found to be simple, and effective.[2] Using this technique, each parameter returns the log ratio of the likelihood that the token is nasalized to the likelihood that the token is non-nasalized. Likelihoods are established by using normalized histograms of the measurements, based on the training data provided to the systems. Bin widths of the histograms were manually set to ensure that the distributions would be reasonably shaped.

The Experiment

As an initial evaluation, the detection system was tested with the utterances of the database. There were a total of 685 nasalized vowels and 500 non-nasalized vowels.

In order to approximate a speaker-independent task given the limited amount of available data, the system was evaluated using a rotational procedure. In each step, system was trained on the data from five of the six speakers in the database, and was tested on the data from the sixth speaker.

Six measures from the acoustic study were incorporated into the detection system. Each measure was taken from one of the three vowel subregions and was usually a maximum or minimum of the three values. The six measures were:

1. *Center of Mass.* The average center of mass in the middle subregion.

2. *Standard Deviation.* The maximum value of the average standard deviation.

3. *Maximum Resonance Percentage.* The maximum percentage of the time there is an extra resonance in the low frequency region.

4. *Minimum Resonance Percentage.* The minimum percentage of the time there is an extra resonance in the low frequency region.

5. *Maximum Resonance Dip.* The maximum value of the average dip between the first resonance and the extra resonance.

6. *Minimum Resonance Difference.* The minimum value of the average difference between the first resonance and the extra resonance.

Results and Discussion

Using the circular evaluation procedure described earlier, we performed the recognition experiments on the speech of all of the speakers in our database. The results of our experiments are summarized in Table 1. While the results vary as a function of the recognition experiments, there are several trends that are worth noting. Across all speakers, an average nasalization detection rate of 74% was obtained. In all but one of the experiments summarized in Table 1, the system is better at detecting nasalized vowels than non-nasalized ones. This is perhaps a reflection of the emphasis that we have placed on the discovery of acoustic features that characterize nasalized vowels, as opposed to oral vowels. Another striking result is that the system performed significantly better for males than for females. The system recognized high vowels hetter than low vowels, a consequence of the fact that the detection rate for low vowels spoken by females is quite poor.

Evaluation	Detection Rate		
	Nasalised	Non Nasalised	Average
All	81	67	74
Male	83	78	81
Female	66	60	63
High	82	75	79
Low	75	63	69
Male High	82	75	79
Male Low	88	83	85
Female High	74	71	73
Female Low	56	67	61

Table 1: Nasalized Vowel Detection Rates

While 74% correct is better than chance, it still leaves a large number of vowels for which no confident statement may be made about whether they are nasalized or not. This is primarily due to the fact that different speakers nasalize to varying degrees, and that the acoustic characteristics of nasalized vowels depend somewhat on the characteristics of the vowel. Thus, by attempting to operate in a speaker-independent mode, the individual distributions are being smeared.[3]

The findings of our acoustic study, as well those by other researchers, have shown that the extra resonance tends to be more *distinct* in low vowels than in high vowels, and speakers tend to nasalize low vowels more readily than high vowels. Thus, one would expect to be able to detect nasalization more successfully in low vowels, as confirmed by our recognition experiment for male speakers.

The poor detection scores for female speech is perhaps also understandable, since there is often an extra low resonance in the sonorant regions. Given that the extra resonance is a major acoustic difference between nasalized and non-nasalized vowels, it is natural to expect system performance to deteriorate when the low resonance is present in the speech signal irrespective of nasality. It is interesting to note that for female speakers, the system was able to identify nasalization in high vowels better than in low vowels. Since the low resonance of female speakers is always below the first formant, high vowels which have a nasal resonance *above* the first formant are uniquely nasal, and so, may be identified correctly.

Finally, it should be noted that our detection algorithms made no use of information regarding the presence of an adjacent nasal murmur. In a separate study, we have found that nasal murmurs can be detected with high reliability [2]. It is very likely that recognition results can be improved substantially when this further source of knowledge is incorporated. While the results are only moderately successful, we were nevertheless comforted by the fact that human listeners do not perceive nasality too well when presented only with an isolated vowel. In a perceptual experiment that we conducted using a subset of the database, it was found that human performance is about the same as that of our recognition system on the same data.

The system was also clearly hampered by our definition of a nasalized vowel since, as has been pointed out, vowels may be nasalized in any context. If the task of the system were to detect truly nasalized vowels rather than vowels adjacent to a nasal consonant, the performance would probably improve.

SUMMARY

The detection of nasalization in vowels is hoth an important and a difficult problem in phonetic recognition. It is important because of the potential benefits that one can derive from their successful detection, as discussed earlier. For the purposes of assisting the detection of nasal consonants, it is difficult because nasalized vowels are not distinguished phonemically from oral vowels in American English. Thus speakers are free to nasalize vowels to various degrees, and may nasalize a vowel in any phonetic context.

In this study, we have established a set of acoustic measures that characterize different aspects of the average spectra of nasalized vowels. Algorithms for extracting these measures automatically from the acoustic signal have been developed, and a set of recognition experiments were performed using these measurements. Our results suggest that vowel nasalization can be detected with moderate success, although the recognition experiments should he validated by using a large amount of new data from unknown speakers. In addition, information regarding the presence of an adjacent nasal murmur may also prove to be helpful.

NOTES

1. The vowels produced by female speakers can have a low resonance in any context. This property is due to breathiness more than nasalization.

2. Gaussian decision making techniques and binary tree classifiers were also examined.

3. Some earlier speaker-dependent experiments obtained detection rates over 10% better than those reported here.

REFERENCES

[1] Fujimura, O., Lindqvist, J., "Sweep-Tone Measurements of the Vocal Tract Characteristics", *Journal of the Acoustical Society of America*, Vol. 49, No. 2, pp. 541-558, 1971.

[2] Glass, J.R., Zue, V.W., "An acoustic study of nasal consonants in American English", Paper presented at the 108th meeting of the Acoustical Society of America, Minneapolis, MN, 1984.

[3] Glass, J.R., "Nasal Consonants and Nasalized Vowels: An Acoustic Study and Recognition Experiment", S.M. Thesis, Massachusetts Institute of Technology, 1984.

[4] Hattori, S., Yamamoto, K., Fujimura, O., "Nasalization of Vowels in Relation to Nasals", *Journal of the Acoustical Society of America* Vol. 30, No. 4, pp. 267-274, 1958.

[5] Hawkins, S., Stevens, K.N., "A cross-language study of the perception of nasal vowels", Paper presented at the 105th meeting of the Acoustical Society of America, Cincinatti, Ohio, 1983.

[6] Rabiner, L.R., Schafer, R.W., *Digital Processing of Speech Signals*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1978.

[7] Roads, C., "A Report on SPIRE: An Interactive Audio Processing Environment", *Computer Music Journal*, Vol. 7, No. 2., Summer 1983.

The Effect of Speech Rate on the Application of Low-Level Phonological Rules in American English

Kimberly Moore & Victor W. Zue
Department of Electrical Engineering and Computer Science, and
Research Laboratory of Electronics
Massachusetts Institute of Technology

This paper was presented on April 10, 1985, at the 109th Meeting of the Acoustical Society of America in Austin, Texas. The research was supported by the Office of Naval Research u .der contract N00014-82-K-0727.

The Effect of Speech Rate on the Application of Low-Level Phonological Rules in American English

Introduction

This paper is concerned with the effect of speech rate on the acoustic characteristics of speech sounds. The primary goals of the study are to improve our understanding of the changes in the continuous speech signal as a function of speech rate, thereby leading to models that will account for these changes. Our hope is also that a fundamental understanding of this sort will benefit efforts in developing speech synthesis and recognition systems that must deal with variability in speech rate. We should note at the onset that the effects of speech rate, from articulatory, acoustic, and perceptual standpoints, have been studied by many other researchers. Most of these studies, however, either have dealt with some of the more global properties, such as the frequency of pause insertion, or have not been concerned with natural continuous speech. Our focus is somewhat different. Specifically, we are interested in investigating changes in the relative frequency of application of certain low-level phonological rules, such as flapping and palatalization. Furthermore, we are interested in quantifying changes in some of the *segmental* cues as a function of speech rate. Our analysis of the data is not complete. This paper should be viewed as a progress report.

Data Collection and Analysis

The corpus that we used consists of a short paragraph containing 47 words in 4 sentences. It is especially designed such that many low-level phonological rules may be applied at most of the word boundaries. In fact, a speaker has the option of applying one or more rules at 35 out of the possible 43 word boundaries. The rules include palatalization, glottalization before vowels, gemination, and alveolar flapping. Some examples of these rules are shown below.

Palatalization	*Could you ...*
Gemination	*Advertisements seem ...*
Flapping	*What ever ...*
Schwa Devoicing	... *to be* ...
Glottal Stop Insertion	.. *such ads.*

The paragraph was recorded by four speakers, two male and two female, at three different rates: fast, normal, and slow. Since it is difficult to control the absolute speech rate from speaker to speaker, we decided to solicit from all speakers fast and slow readings relative to their normal speech rate. The recording procedures were as follows. First, the speakers were asked to read the paragraph several times at their normal rate. The average duration of the reading was measured with a digital clock. For the slow reading, the clock was set to twice the time of the normal reading, and the speakers were asked to complete the reading in the allotted time. For the fast reading, we had originally hoped to gather data at twice the normal rate. However, it became clear early on that people have extreme difficulty speaking at twice their usual rate without significantly affecting intelligibility. As a result, we modified our procedure and asked the speakers to complete the reading in three-fourths the time of their normal reading. A minimum of two readings for each rate was recorded. All in all, our database contains 48 paragraphs, a total of slightly over 2,200 word tokens.

Figure 1 summarizes the actual speech rate measured in terms of the number of syllables per second of speech, for each experimental condition and for each of the speakers. We see that for both the fast and the slow speaking conditions, speakers can indeed produce speech at the desired rate. The average number of syllables per second is 2.4, 4.8, and 6.0, for the slow, normal, and fast reading conditions, respectively.

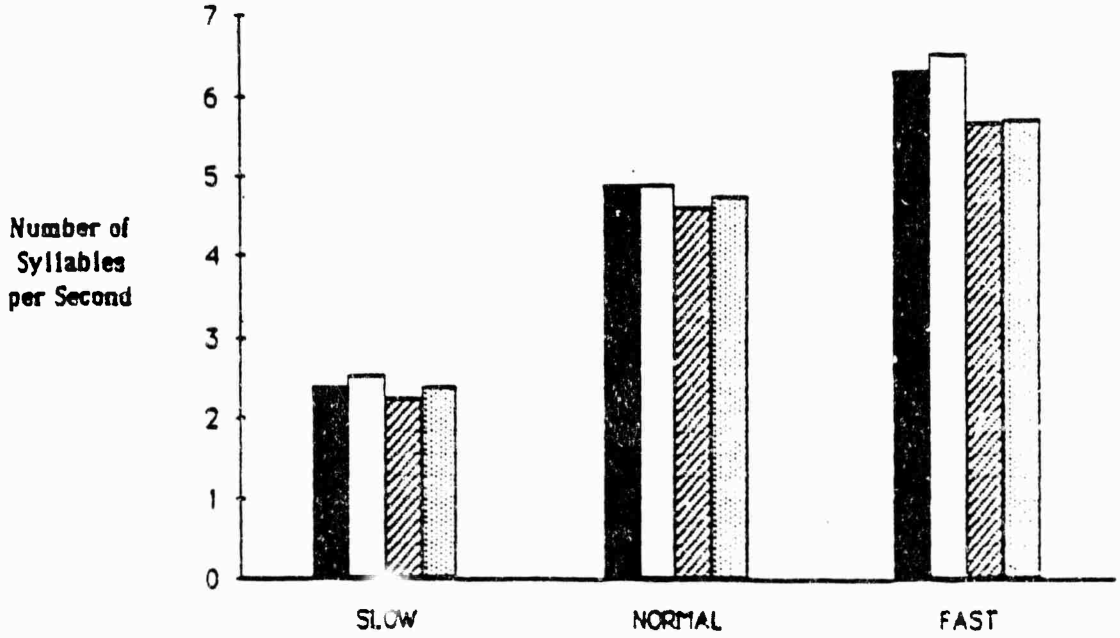


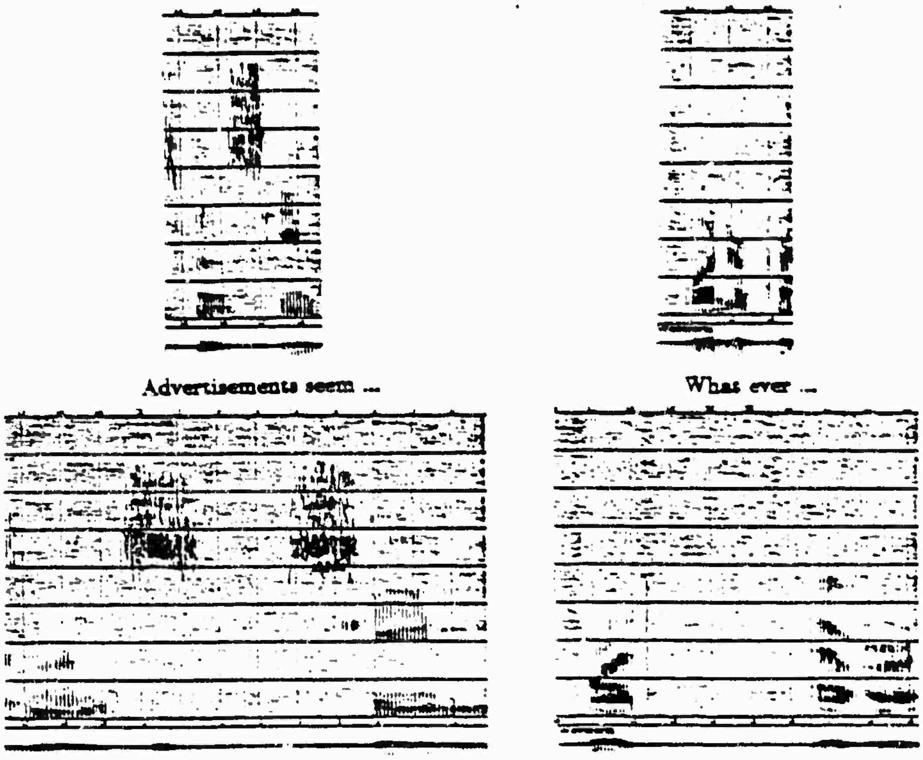
Figure 1: *Average Speaking Rate for the Three Conditions*

All the recorded utterances were digitized at 16 kHz using the *Spire* facility developed at MIT, and digital spectrograms for the utterances were made. Phonetic transcriptions for the utterances were obtained independently by the two experimenters. This was done both by listening to the utterances and by visually examining the spectrograms and other relevant displays. The experimenters discussed the differences, which were often minor, and a consensus was reached. This transcription was then time-aligned with the speech waveform by hand. Most of the data was subsequently analyzed using *SpireX*, a data analysis and statistics gathering program on our Lisp Machine workstations.

Word Boundary Effects

As explained earlier, we designed the short paragraph such that low-level phonological rules can be applied at many of the word boundaries. The analysis of our data indicates that there is a general tendency for the frequency of rule application to be correlated with speech rate.

Figure 2 gives some examples. In the left-hand column, we compare the spectrogram of a portion of the phrase "advertisements seem" spoken by a male speaker at the slow and fast rates. In the top panel, the two /s/ sounds are geminated, whereas in the bottom panel, a long pause is inserted between the two /s/'s. In the right-hand column, we compare fast and slow readings of the phrase "what ever." In this case, the word final /t/ is turned into a flap for the fast condition, whereas a pause and a glottal stop are inserted following a weak release for the slow condition.



Advertisements seem ...

What ever ...

Figure 2: *Spectrograms—Slow vs. Fast Rate*

Our findings for the frequency of application of low-level phonological rules are summarized in Figure 3.

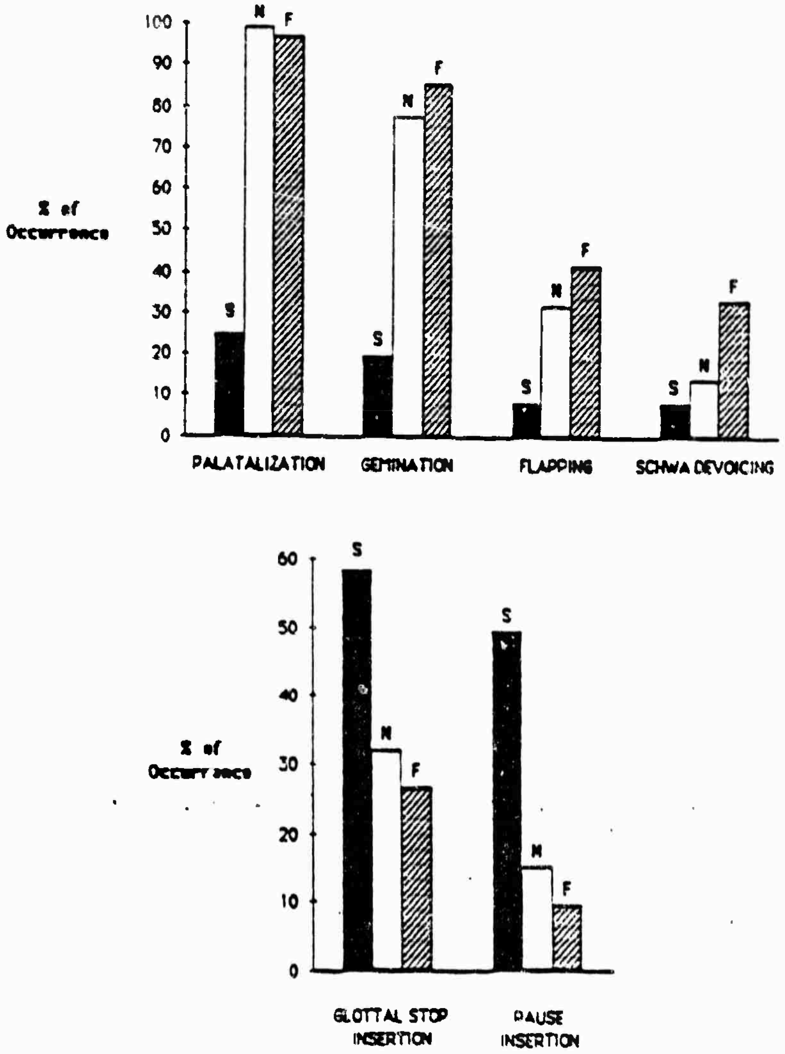


Figure 3: *Frequency of Rule Occurrence*

The top graph describes those rules that occur more frequently as the speech rate increases. We see that for palatalization and gemination, there is a dramatic increase in the frequency of rule application from slow to fast. In fact, these two rules are often applied as frequently in normal speech as they are in fast speech. Flapping and schwa devoicing, on the other hand, do not occur nearly as frequently as the other two. For these two rules, the application at the fast rate varies from speaker to speaker. Schwa devoicing, for example, does not occur in significant number until the fast condition, and as such is dominated by two of the four speakers.

The bottom graph shows those rules whose frequency of occurrence is negatively correlated with the speech rate. At the slow rate, three of the speakers inserted pauses at almost all the word boundaries, and the speech is read as if the sentences were strings of isolated words concatenated together. When pauses are inserted before a word that starts with a vowel, a glottal stop is often observed. Again, the frequency of glottal stop insertion is about the same for normal and fast speech.

Segmental Effects

We now turn our attention to the second issue, namely the segmental changes due to changes in speech rate. For this presentation, we will limit our discussion to the durational changes for various speech sounds. In general, the segmental durations increase as the speech rate decreases.

This is illustrated in the next figure, which compares the vowel durations for the three experimental conditions. We see that the histogram for the fast rate is similar to that of the normal rate. The decrease in vowel duration is small, presumably due to to the incompressibility of segment durations as suggested by Klatt and others. Focusing now on the slow condition, we see that vowels can be lengthened significantly when speech rate is reduced.

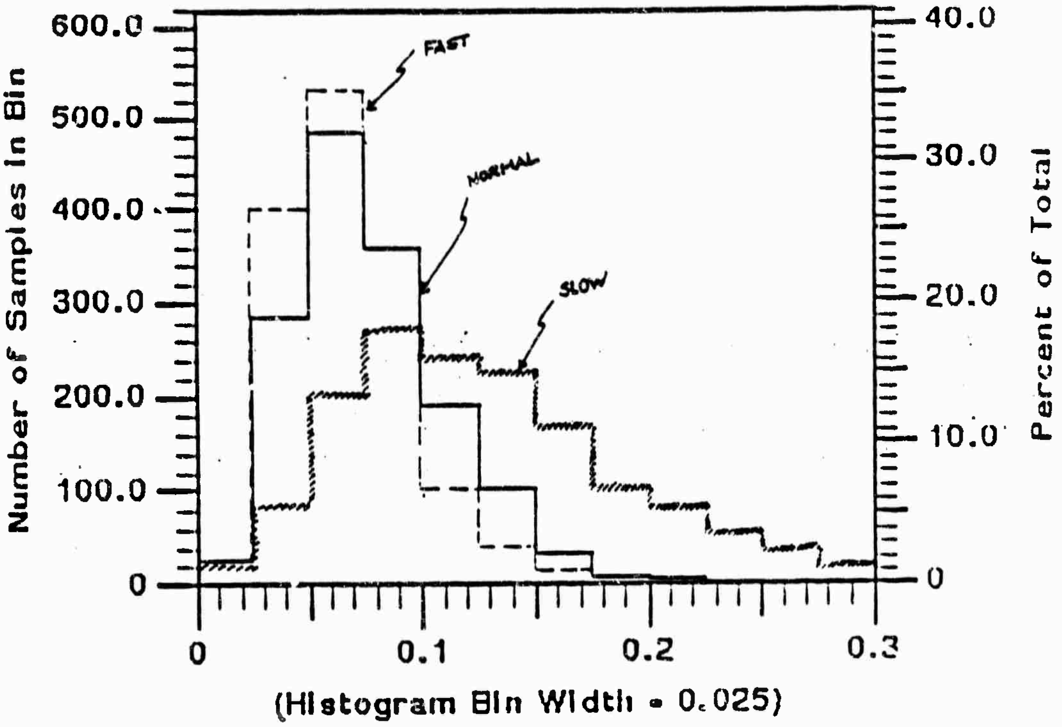


Figure 4: *Histogram of Vowel Duration for the Three Conditions*

While the general trends suggested in this figure hold true for all the speech sounds, the amount of change varies from sound to sound. Figure 5 summarizes for several classes of speech sounds the percentage change in the average duration, as compared to the normal condition, for the fast and slow conditions.

As can be seen from this figure, the amount of durational increase for the slow condition is greater for vowels and nasals, and smaller for stops and fricatives. This is presumably due to the fact that the airflow is greater for the turbulent sounds, making them harder to sustain. In contrast, the decrease in duration for the fast condition is around 14%.

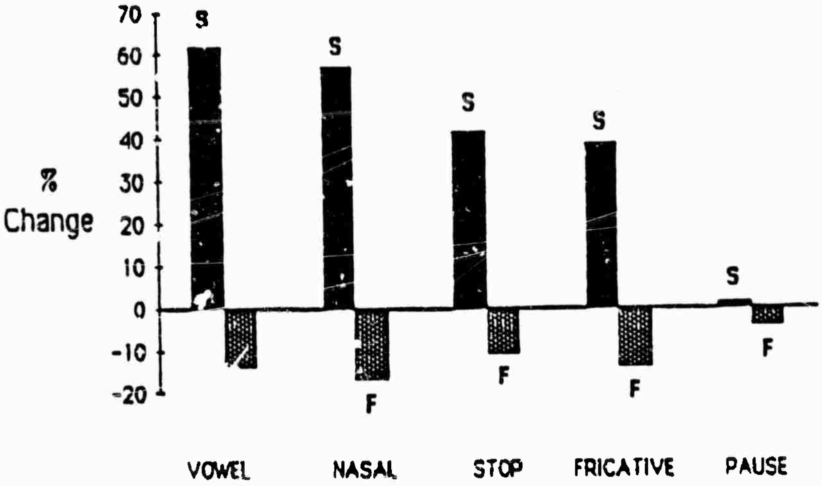


Figure 5: *Percentage Change in Segmental Directions (Relative to the Normal Rate)*

The amount of change in segmental duration is significantly less than what the the actual speech rate would predict. This is due to the fact that the predominance of the overall durational changes can be accounted fc by pauses. The fast condition contains one-third fewer pauses than the normal condition. Since the pauses are considerably longer than speech sounds, a decrease in the proportion of pauses further increases the overall speech rate. For the slow condition, there are over three times as many pauses inserted between words. As a result, the overall speech rate is further reduced. While the number of pauses increases significantly from the fast to slow conditions this figure shows that the average duration of a pause remains relatively unchanged for all three conditions.

Summary

In summary, our analysis indicates that for most low-level phonological rules, their relative frequency of application increases when the speech rate is increased from the speaker's normal rate. In contrast, speakers often adopt different strategies for slowing down their speech, including the insertion of pauses and the release of word-final stops, such that the frequency of application of the rules varies from speaker to speaker.

When the speech rate is faster than normal, our results indicate that the segmental durations are decreased almost uniformly. This is accompanied by a reduction in the number of pauses in the sentences. When the speech rate is slower than normal, the increase in segmental duration appears to vary from sound to sound. Our result is in agreement with previous work by Goldman-Eisler, Huggins, Grosjean, and others, that in slow speech there is a sharp increase in the number of pauses in a sentence, with each word taking on the appearance of an isolated word. While the number of pauses varies as a function of speech rate, the average duration of a pause remains unchanged.

Paper presented at Speech Tech 85 Conference, New York, NY, April 24, 1985.

The MIT *Spire* System

Victor W. Zue and D. Scott Cyphers
Department of Electrical Engineering and Computer Science and
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139

Introduction

In many areas of speech research, ranging from speech analysis and synthesis to recognition, researchers are often faced with a common set of analysis procedures. Specifically, there is often the need to:

- Record and digitize utterances,
- Create various attributes of the speech signal, and
- Display and perform interactive measurements on these various attributes.

The ease with which one can perform these tasks greatly facilitates the gathering of information and the corresponding improvement of our speech knowledge. *Spire* (*Speech and Phonetics Interactive Research Environment*) represents our attempt to provide the answer to such research needs at MIT.

Spire was originally designed and implemented by David W. Shipman [Shipman, 1982]. Since 1983, however, the system has undergone considerable modification by the second author of this paper, DSC. It is an evolving system that is still being actively improved. This paper serves as a progress report on the present status of *Spire*.

Hardware Requirements

The speech workstation that we have developed centers around a Lisp Machine, originally developed at the M.I.T. Artificial Intelligence Laboratory. It is specifically designed for the efficient execution of Lisp, a symbolic programming language widely used in the artificial intelligence community. The current version of *Spire* runs on a Symbolics 3600 or 3670 Lisp Machine. The Lisp Machine has 2 Mbytes of main memory and a 1 Gbytes address space, a 474 Mbyte disk, and a *Spire* Unibus interface. The Lisp Machine's high-resolution graphics console and hand-held pointer allow development of extremely convenient user interfaces.

We have augmented the standard configuration of the Lisp Machine with an FPS-100E array processor (up to 4 Mflops), which handles essentially all the computationally intensive numeric processes. The work station also includes a DSC-200/240 A/D and D/A converter and audio amplifier, a Versatec V-80 electrostatic printer/plotter, and assorted audio equipment such as a microphone, a set of headphones, and a tape recorder. The Lisp Machine work stations are connected to one another and to central file servers via a packet-switched local area network.

System Description

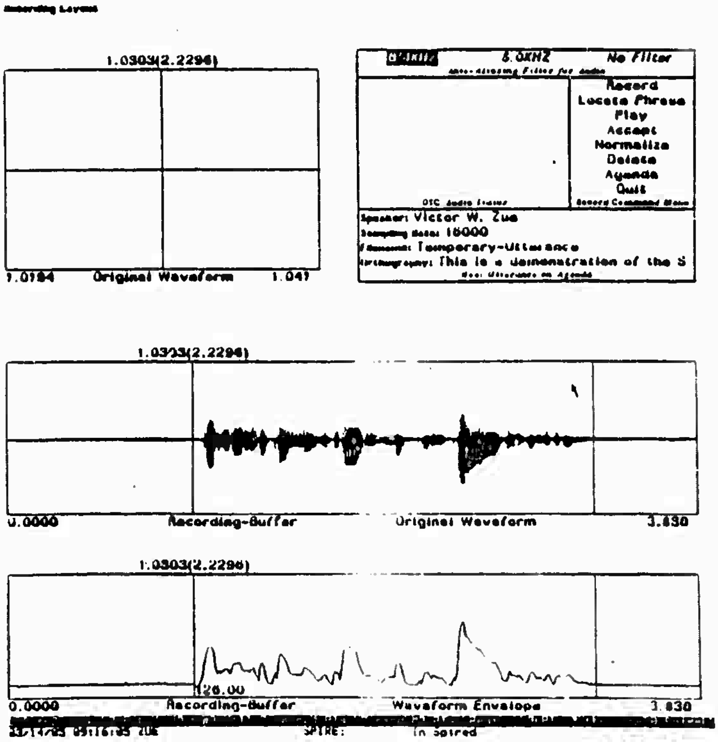
Spire organizes an utterance as a collection of *attributes*. The attributes may be symbolic (e.g. phonetic transcription), or they may be numeric (e.g. RMS amplitude). Some of the attributes are one dimensional (e.g. speech waveform), while others are multi-dimensional (e.g. short-time spectra). *Spire* has knowledge of the properties of the attributes, as well as how they are computed. As a result, attributes can be displayed conveniently, and they can also be used to compute new attributes.

Spire displays are organised in the form of *layouts*. A layout is a collection of displays that the user can compose freely to suit his/her research needs. Layouts that are used frequently can be pre-defined and saved for future usage. For example, the recording layout and the transcription layout are provided by *Spire*, since they are almost always needed by a user, and serve as example for beginners. Many of the commands in *Spire* are executed by means of the hand-held mouse pointer. The mouse can be used to configure a layout, to play back a section of the utterance, edit waveforms, examine data values, alter display options, and perform other functions.

For the remainder of this section, we will give some examples of the operation and capabilities of *Spire*. For a detailed description of *Spire*, see Shipman [1982], and Cyphers [1985].

Recording

Figure 1 shows the recording layout of *Spire*. The default sampling rate is either 16 kHz or 20 kHz, although it can be as high as 70 kHz. Appropriate anti-aliasing filters can be selected by the user. Information about the talker, sampling rate, filename, and the orthographic transcription can all be changed easily with a click of the mouse. Alternatively, an agenda file can be set up to sequentially change these parameters. This latter option is particularly useful for bulk data input when a list of the utterances



Speaker: Victor W. Zue
Sampling Rate: 16000
Filename: Temporary-Utterance
Orthography: This is a demonstration of the S
Next Utterance on Agenda

Figure 1: The *Spire* recording layout.

already exists on-line.

Currently, *Spire* can accept an essentially unlimited amount of speech at a stretch. An automatic end-point detector attempts to locate the utterance. The user can listen to the located utterance, modify the endpoints if necessary, and accept the utterance into the database, all with several clicks of a mouse button.

Transcription

Figure 2 shows the transcription layout of *Spire*. This layout is used to enter the phonetic transcription, and time-align it (or the orthographic transcription) with the speech waveform. For phonetic alignment, the region of the waveform bounded by the cursor (shown as the solid vertical line in the spectrogram and waveform displays) and the marker (shown as the dotted line in the same displays) can be associated with a phonetic symbol by a click of a mouse button. Using this layout, an experienced acoustic phonetician can align a two second utterance in about 5 minutes.

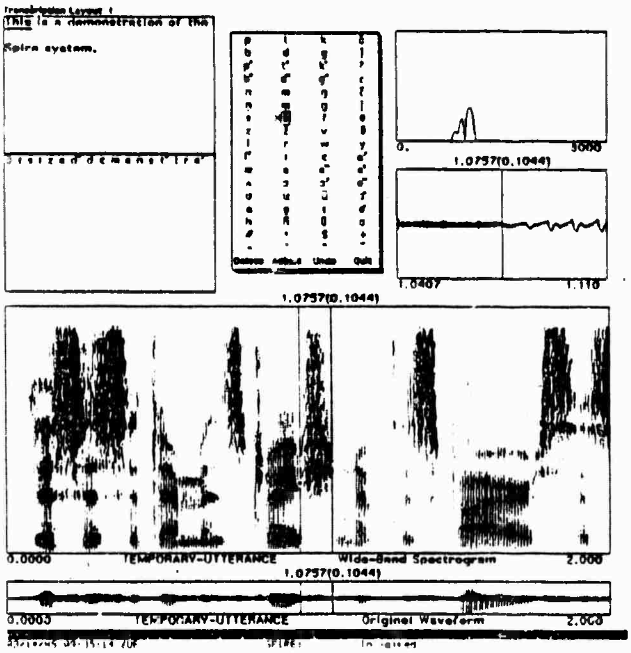
While manual time-alignment using *Spire* is quite efficient, it nevertheless requires the expertise of a small group of experts. As a result, the amount of data that can be collected and aligned is greatly limited. In addition, phonetic alignment is often subjective, thus leading to inconsistencies among transcribers. The tedious nature of the task also tends to introduce human errors. We have recently developed a semi-automatic system, extending the basic capabilities of *Spire*, to perform the time alignment. The results of our preliminary evaluation have been encouraging. For a description of the alignmert system, see Leung and Zue [1984] and Leung [1985].

Other Features

One of the most important features of *Spire* is that a user can compose his/her own layouts for the specific research needs. Figure 3 shows an example of such a layout. The figure displays the wideband spectrogram of the utterance, the zero crossing rate, the original waveform, the orthographic as well as phonetic transcriptions, and the short-time, narrow band and LPC spectra. This layout illustrates some of the interactive features of *Spire*. First, display parameters can be changed by the user at will. Thus for example, the zero-crossing rate is displayed on the same time scale as the wideband spectrogram. Second, all displays are time-synchronized. Moving the cursor in one display will cause the other displays to change accordingly. Third, displays can be overlaid, and the overlay parameters can be changed as well. For example, the narrowband spectra are overlaid with the LPC spectra, and the LPC spectra are displayed with a different thickness for distinction.

Spire also has the capability of generating high-quality digital spectrograms. The output device has a resolution of 200 points per inch. Figure 4 gives an example of a digital spectrogram.

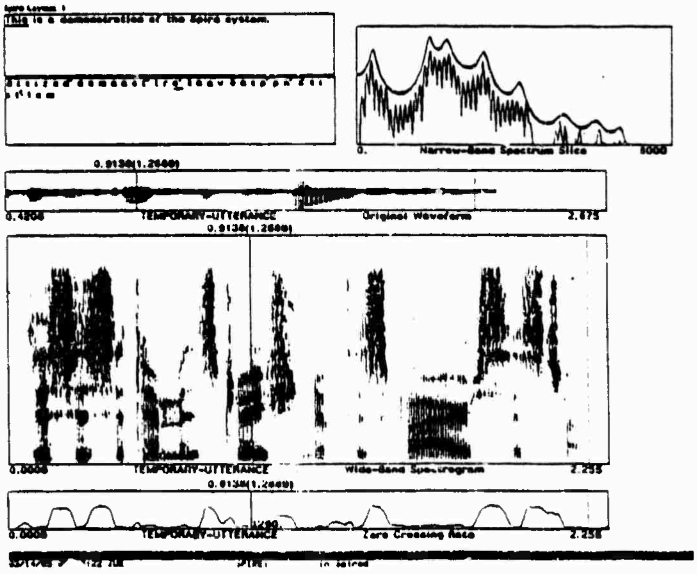
Spire was designed with two general goals in mind. First, it provides an extremely interactive environment and a basic set of capabilities such that speech scientists, with little or no programming experience, are able to collect and analyze speech data. Second, it provides a framework for users to conveniently develop



Transcription Layout 1

This is a demonstration of the Spire system.

Figure 2: The *Spire* transcription layout.



Spire Layout 1

This is a demonstration of the Spire system.

Figure 3: A typical *Spire* screen layout.

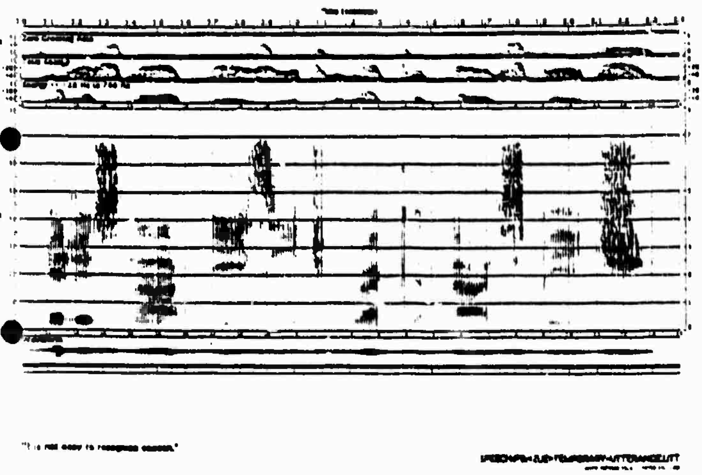


Figure 4: A digital spectrogram produced with *Spire*

systems within *Spire*, i.e. to add new attributes, to suit their own research needs. Currently the default version of *Spire* contains approximately 40 attributes of the speech signal, whereas a customized version can compute any number of attributes. Some of the customized *Spire* versions in our research group have as many as three hundred attributes.

Summary

The development of the *Spire* system was a 5 man-year effort, spread over a period of three years. Our goal is to create a research environment that is easy to use, and thus increase the amount of data that a speech scientist can examine, leading to an increase in our speech knowledge. It has played an important role in advancing our understanding of the acoustic properties of speech sounds.

While *Spire* is still being actively improved, we are eager to share our development results with other researchers who may find such a system useful. In fact, the system configuration has been duplicated, and the software acquired, by many research laboratories and universities outside of MIT. Those who are interested should contact the MIT Patent Office directly for licensing procedures.

Acknowledgement

This research is supported by DARPA, as monitored through the Office of Naval Research under contract N00014-82-K-0727, and by the System Development Foundation. The *Spire* system was originally designed and implemented by David W. Shipman. David Kaufman has also contributed to various phases of its development.

References

1. Cyphers, D. S. (1985), "*Spire*: A speech research tool," S.M. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

2. Leung, H.C. and Zue, V.W. (1984), "A procedure for automatic alignment of phonetic transcription with continuous speech," Proceedings of the 1984 International Conference on Acoustics, Speech, and Signal Processing, San Diego, CA, 2.7.1-4.

3. Leung H. C. (1985), "A procedure for automatic alignment of phonetic transcription with continuous speech," S.M. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

4. Shipman, D.W. (1982), "Development of Speech Research Software on the M.I.T Lisp Machine," J. Acoust. Soc. Am., 71, Suppl. 1, S103.

Paper presented June 24, 1985, at IASTED Robotics & Automation Conference, Lugano, Switzerla

Analysis and Recognition of Nasal Consonants in American English

James R. Glass and Victor W. Zue
Department of Electrical Engineering and Computer Science and
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139

ABSTRACT

This paper deals with the analysis and recognition of nasal consonants, /m, n, ŋ/, in American English. The acoustic analysis was performed on a database of over 1,200 words excised from a carrier phrase and spoken by six speakers, three male and three female. Across all speakers, we found that nasal murmurs are characterized primarily by a lower spectral amplitude than the adjacent vowels, and by the presence of a dominant, low-frequency peak in the short-time spectra. Automatic procedures were then devised to reliably extract acoustic attributes that reflect these characteristics. Finally, recognition experiments were performed to test the validity of these attributes. Our results, based on 600 sentences spoken by 60 speakers, shows that nasal consonants can be distinguished from the imposters with an accuracy of 83.5%.

INTRODUCTION

This paper deals with the analysis and recognition of nasal consonants, /m, n, ŋ/, in American English. Nasal consonants are produced with a closure in the oral cavity. By lowering the velum, airflow is directed through the nasal cavity and eventually radiated from the nostrils. The transfer function for nasal production contains poles as well as zeros, the latter a consequence of the fact that the vocal cavity serves as a side branch. Thus the power spectrum of nasal consonants shows spectral prominences as well as spectral notches [3].

In American English, nasal consonants appear quite frequently. (They have a combined frequency of occurrence of about 11% [9].) Nasals can be singly attached to a vowel, or they can form a cluster with other consonants. Nasal consonants are difficult to recognize for several reasons. First, the characteristics during oral closure, often referred to as the nasal murmur, differ significantly from speaker to speaker because of differences in the size and shape of the nasal and sinus cavities. Second, a nasal murmur can also be affected drastically by the phonetic environment. In some cases, the nasal murmur is almost entirely absent (as in "camp") [5]. In this case the presence of the nasal consonant lies almost entirely in the degree of nasalization in the adjacent vowel. Finally, the complex production mechanism makes them difficult to characterize acoustically.

The goal of our research is twofold: (1) We are interested in discovering speaker-independent acoustic cues for nasal murmurs, and (2) We would like to test the effectiveness of these cues in actual recognition experiments where the nasal consonants are to be distinguished from imposters. While the acoustic characteristics of nasal murmurs have been studied extensively in the past [2], [3], these results are not directly applicable to speech recognition. In some cases, the acoustic analyses were not based on a sufficient amount of data. Most of these studies were primarily concerned with isolated words. In addition, the acoustic features do not always lend themselves to automatic extraction and subsequent computer recognition.

We must emphasize that analysis of the nasal murmur will only provide partial information regarding the presence of nasal consonants. An integral part of our study deals with the analysis of vowel nasalization [4], which will not be dealt with in this paper.

DATABASE DESCRIPTION

The acoustic analysis of nasal consonants was made from a database specifically created for this study. The corpus was based on a set of some two hundred carefully chosen words that contain nasal consonants in many different phonetic contexts. For example, the nasal consonant may appear prevocalically (as in "mitt"), medially (as in "simmer"), postvocalically (as in "dim"), and in clusters (as in "snow" and "think"). In addition, some words containing *imposters*, i.e. other consonants that are acoustically similar to the nasal consonants (as in "demise/devise"), are also included. Many of the words form minimal pairs (as in "sin/sing/sick/sink/sinking") so that the effect of the phonetic context can be isolated and subtle acoustic differences identified.

Once the corpus had been designed the database was created. Three male and three female speakers each read the words of the corpus which were embedded in a carrier phrase. This resulted in a database of slightly over 1,200 words. All utterances were digitized at 16 kHz and the phonetic transcriptions were manually time aligned with the waveforms. Temporal measurements were made from the time-aligned transcription. while spectral measurements were made from a cepstrally smoothed short-time spectrum.

ACOUSTIC ANALYSIS

Apart from the manual alignment of the phonetic transcription with the speech waveform during data preparation, all other measurements and analyses were performed without human intervention. This automatic procedure permits immediate implementation for nasal recognition once an acoustic attribute has been shown to be robust. Another advantage of automatic analysis is that a large amount of data can be analyzed in a reasonable amount of time. Thus the size of the database was limited primarily by the time it took to complete the time alignment. Automatic analysis must be done carefully however, or one runs the risk of adding measurement noise into the distributions of the data.

Figure 1 shows the spectrograms of the words "simmer", "sinner", and "singer", spoken by a male speaker. These spectrograms serve to illustrate some of the qualitative features common to all nasal consonants. We see that the nasal murmur typically has lower energy than the adjacent vowels. It is also delineated from the vowels by a sharp spectral discontinuity. In addition, it is characterized by the presence of a low frequency spectral peak. There are, however, other sounds with acoustic characteristics similar to those of nasal murmurs. Some of these sounds, a voice-bar (referring to the closure portion of voiced stops) and a prevocalic /l/, are shown in Figure 2.

The acoustic analysis is focused on quantifying the features shown in the spectrograms. We started by measuring the duration of nasal

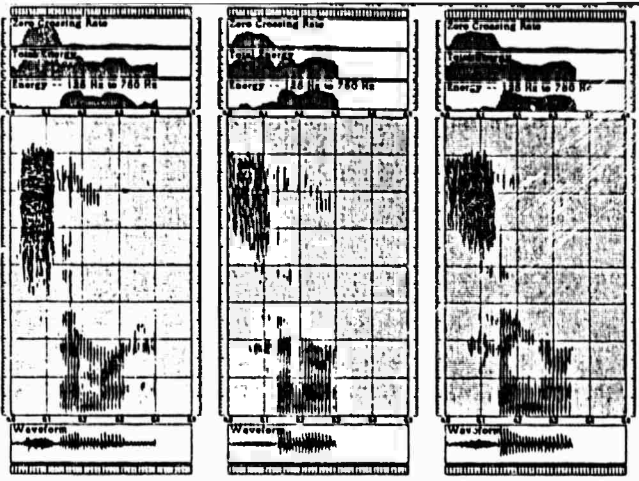


Figure 1: Spectrogram of the words "simmer", "sinner" and "singer", spoken by a male speaker.

murmurs. In agreement with previous studies [8] we found that the duration of the nasal murmur is strongly influenced by phonetic context. For example, the nasal murmur is shortened when it is in a cluster with a voiceless consonant, and is lengthened when it is in a cluster with a voiced consonant. This result was found to be true for both word-initial clusters, as in "smack", and word-final clusters, as in "can't". Although these differences are fairly robust, our ability to utilize such information in speech recognition must depend on knowledge of their exact context and the speaking rate.

We next investigated the energy in the nasal murmur relative to the adjacent vowels. This energy difference is determined by subtracting the average total energy in the nasal murmur from the average total energy in the adjacent vowels. For the tokens in the database, this energy difference was almost always positive, implying that the nasal murmur is consistently weaker than an adjacent sonorant.

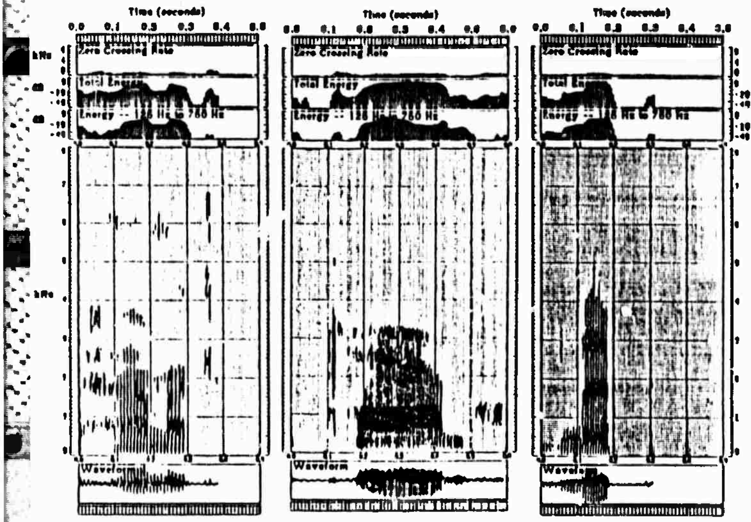


Figure 2: Spectrogram of the words "ha mmock", "cab" and "lip", spoken by a male speaker.

As illustrated in figure 3, nasal consonants in a medial position have a slightly smaller energy difference than nasal consonants in other contexts. This is probably due to the fact that medial nasals have a sustained level of energy when surrounded by vowels. In contrast, energy during nasal murmurs tends to rise gradually in prevocalic positions, and taper off in postvocalic positions, resulting in a lower average value. This observation is reinforced by studies of the nasal murmur stability, which indicate that the energy of medial nasals is quite steady.

Figure 3 also compares the value of the energy difference of the nasal consonants to similar sounds such as liquids and glides, and voice bars, overlap in the distributions, it is clear that, on the average, nasal consonants have a greater decrease in energy than liquids and glides, and have a smaller difference than voice bars. Thus, from a speech recognition point of view, this measurement may help to distinguish nasals from liquids and glides.

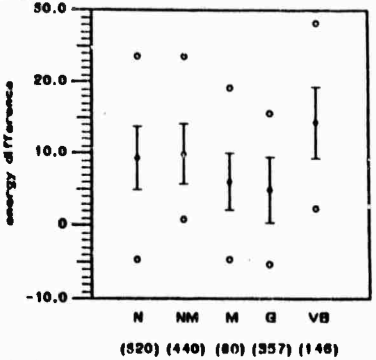


Figure 3: Statistical Summary of the Average Energy for All Nasals (N), Non-Medial Nasals (NM), Medial Nasals (M), Liquids and Glides (G), and Voice Bars (VB).

(The average value is indicated by a filled circle. The vertical lines indicate one standard deviation and the open circles display maximum and minimum values. The number of samples in each context is indicated below the display.)

Spectral analysis of the nasal consonants was based on a cepstrally smoothed spectrum created from the pre-emphasized waveform. The spectra were all normalized with respect to total energy to eliminate offsets. During analysis we also restricted ourselves to analyzing the spectra of one rpeaker at a time since we found that the spectra, primarily at frequencies above 1000 Hz, were highly speaker dependent. This dependency is probably due to the fact that the size of the nasal and sinus cavities can vary greatly from speaker to speaker.

Statistics were gathered by collecting multiple spectra, computed every 10 ms, from all of the nasal murmurs. Although the averaging procedure can potentially smear useful information, it serve the purpose of revealing general trends across all phonetic contexts. Figure 4 shows examples of the average spectra for the three nasal consonants for a male speaker. Although subtle differences could be detected among the three nasal consonants for any given speaker, such difference are overshadowed by their similarities. This observation has been made previously by Fujimura, who also found little differences among the spectra of the three nasal murmurs [3]. Furthermore, we found that the spectral shape of the nasal murmur was relatively unaffected by phonetic context.

A more quantitative analysis verified these general observations. Specifically, we found that the nasal murmur spectra were characterized by a low frequency resonance which dominated the spectrum. This low frequency spectral peak was nearly always centered between 200 and 350 Hz. The amplitude of this low frequency resonance is quite stable, and is higher for nasals than for semivowels, as shown in Figure 5. Another characteristic of the nasal murmur was an abrupt decrease of energy in the frequencies immediately following the low frequency resonance. Again, this attribute can be effective in distinguishing nasal murmurs from semivowels.

RECOGNITION EXPERIMENTS

After establishing some of the acoustic properties of nasal consonants and developing attributes that capture these properties, preliminary investigations were conducted to evaluate the usefulness of these attributes in speech recognition. In order to be consistent with the ways the acoustic analyses is carried out, we have structured the recognition experiments as a set of *discrimination* tests. Specifically, the nasal detection system is given a test token and a set of training data. The system must then classify the token as either a nasal murmur, or an imposter sound, such

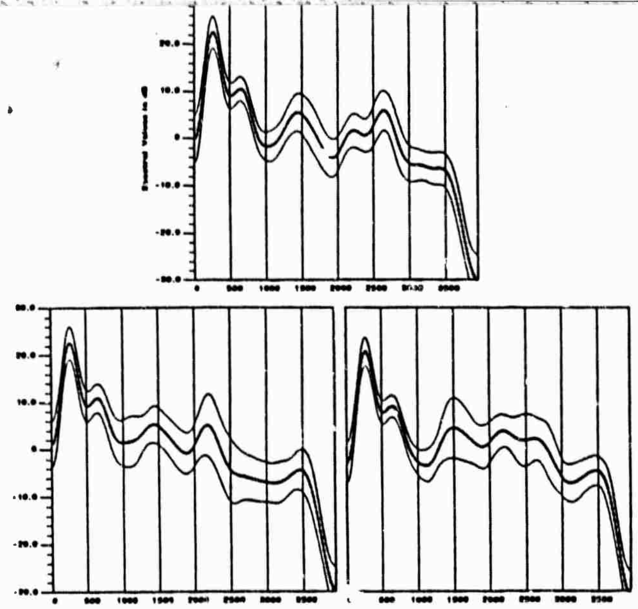


Figure 4: Average Spectrum of the Nasal Consonants /n/ (Top), /m/ (Bottom Left), and /ŋ/ (Bottom Right), for a male speaker. (The average spectrum, shown by the dark line, is surrounded by lines which represent one standard deviation from the mean.)

as a liquid, glide, voice bar or voiced fricative. Throughout these experiments, it is assumed that the boundaries of the murmur are known, and that there is some knowledge of the broad phonetic context. For the experiments described in this paper, no information in adjacent sounds is utilized.

The Strategy

Our acoustic study produced several attributes, each potentially useful in characterizing a certain aspect the the nasal murmur. We have chosen to incorporate the five most robust measurements into detection systems for the task in hand. The five measures are:

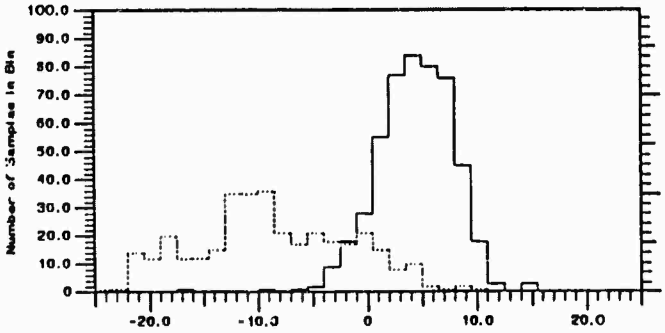


Figure 5: Histograms of the Amplitude of the Low Frequency Resonance Obtained from the Average Spectra for Nasals (Solid Line, 520 Tokens) and Semivowels (Dashed Line, 357 Tokens).

- *Energy:* The difference in average energy between the consonant and the adjacent vowel,
- *Percentage:* The percentage of the time that there was a low frequency resonance centered between 200 and 350 Hz in the consonant,
- *Strength:* The average amplitude of this resonance in the consonant,
- *Drop:* The average energy drop from the low frequency resonance to the frequency region immediately above, and
- *Stability:* The change in low frequency energy throughout the consonant.

Thus, a given token is associated with a set of five values that correspond to the acoustic measurements made on the test token. If we consider the set of values as a vector in a five dimensional space, we are faced with a multi-dimensional decision making problem.

After examining several possible classification procedures [4], we settled on the use of a binary tree classification technique. Decision tree classifiers have been used in a variety of pattern recognition problems and have a number of important advantages over single stage classification techniques, particularly in problems of high dimensionality [1]. Through the use of the decision tree, a complex global decision is made through a series of simpler, local decisions at each level of the tree. This approach is amenable to situations when the decision surface is complex, or when the number of classes to be identified is large. More importantly, the decision tree is a structure which is easy to interpret and can often provide insight into a particular problem [7]. In addition, knowledge obtained through constructing the tree using the learning sample can be augmented by human supervision and guidance during tree growth.

Experiment One

As a first step in validating the usefulness of the acoustic attributes, the recognition task was performed using the utterances of the original database. 520 nasal murmurs and 695 impostor sounds were used in this experiment. In order to approximate a speaker-independent task given the limited amount of available data, the system was evaluated using a rotational procedure. In each step, the system was trained on data from five of the six speakers in the database, and was tested on the data from the sixth speaker.

Earlier investigation reveals that the usefulness of the attributes depends on knowledge of the broad phonetic context. As a result, the data were first divided into three categories based on the broad phonetic context; prevocalic, post-vocalic, and intervocalic. Apart from this preliminary structure, decision trees were grown automatically using the training data. The results of this experiment are shown in Table 1. We see that, for this database, the system produced an average identification rate of 83.6%.

Table 1: Detection Confusions on Small Database

	Output (%)	
Input	Nasal	Impostor
Nasal	83	17
Impostor	16	84

Experiment Two

Although the results of experiment one were encouraging, there is a possibility that these results were a reflection of the fact that the same database was used for system development and system evaluation. As a result, a second database was collected in order to provide a more realistic evaluation of the recognition systems.

The new database contained ten sentences recorded from thirty male and thirty female speakers. In addition to the phonetic environments investigated previously, this second database also contains nasal consonants interacting with vowels and consonants across word boundaries. All in all, this second database provided over 1100 nasals and 2000 impostor tokens from 60 speakers, in a continuous speech environment. Once again, the system was evaluated in a rotational procedure, this time training on data from fifty of the speakers, and tested on the data of the remaining ten speakers.

Using the broader context produced an average identification rate of 83.5%, which is essentially identical to the results of Experiment 1. A breakdown of the result may be found in table 2.

DISCUSSION

The results of the two recognition experiments demonstrate that the automatically extracted acoustic attributes are useful in distinguishing

Table 2: Detection Confusions on Large Database

	Output (%)	
Input	Nasal	Impostor
Nasal	79	21
Impostor	15	85

nasal murmurs from imposters across a large number of male and female speakers. We are encouraged by these results for several reasons. First, closer examination of the experimental results reveals that the recognition performance varies little from speaker to speaker, suggesting that the acoustic attributes are capturing speaker-independent cues. Second, the binary decision trees for different training samples appear to be very similar, i.e., the same attributes are often used at the same node in different trees. The stability of the decision trees is indicative of the fact that the acoustic attributes, as well as the way they are being utilized are quite robust. Finally, one must keep in mind that we have based the recognition of nasal consonants solely on information contained in the nasal murmurs. In many phonetic environments, the nasal murmur is both weak in amplitude and short in duration. As a result, the nasal murmur may not always provide the clearest information regarding the presence of a nasal consonant. By incorporating knowledge of the degree of nasalization in adjacent vowels, better recognition performance can be expected.

By comparing Tables 1 and 2 we see that imposters are identified with similar accuracy, whereas nasals are identified less accurately in the second, larger database. We believe this difference may be due to the fact that the speech data is acoustically more variable in the second database. (Recall that the first database consisted of words excised from a carrier phrase, whereas the second database contains continuous sentences.) In addition, the second experiment utilized proportionally more imposter than nasals. Since the binary tree classifier inherently incorporate *a priori* frequency of occurrence information into the tree structure, it is expected to perform better for the more likely candidates.

Our preliminary examination of the decision tree suggested that different attributes may be effective in different phonetic contexts. The results of the recognition experiments indicate that this is indeed the case. The average recognition scores 85.8%, 80.3%, and 82.5% for the prevocalic, medial, and postvocalic context, respectively. In fact, the decision tree looks quite different for the three contexts. For example, *stability* was found to be the most useful attribute in the medial context, but not very reliable in the postvocalic context. This is presumably due to the fact that the low frequency energy is higher and more steady in the medial context, as discussed earlier. Note that we have chosen to partition the data in terms of *broad* phonetic context. This approach stems from our belief that such contexts can be established effectively in practical recognition tasks.

Examination of the recognition results indicates that there is a slight difference in performance between male and female speakers. This could be due to differences in vocal tract sizes. Our evaluation procedure, which always test a group of ten all-male or all-female speakers, may have further enhanced this contrast. By incorporating an equal proportion of male and female speakers into the training and test data, slightly better recognition results may be observed.

SUMMARY

In summary, our acoustic analyses resulted in the discovery of some distinct characteristics of nasal murmurs, and we are encouraged by the preliminary results of using these acoustic attributes for nasal recognition. We feel that our study supports the notion that a better understanding of the acoustic properties of speech sounds will lead to improved performance in phonetic recognition.

Future work in this direction includes the characterization and recognition of nasalized vowels, and the utilization of acoustic information in both the vowel and murmur portions to identify the nasal consonants. We also plan to investigate procedures that automatically identify the boundaries between vowels and oral closure.

[This research was supported by the Natural Sciences and Engineering Council of Canada and by the Office of Naval Research under contract N00014-82-K-0727.]

Bibliography

[1] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., *Classification and Regression Trees*, Wadsworth International Group, Belmont, California, 1984.

[2] Fant, G., *Acoustic Theory of Speech Production*, Mouton and Co., 's-Gravenhage, Netherlands, 1960.

[3] Fujimura, O., "Analysis of Nasal Consonants, *Journal of the Acoustical Society of America*, Vol. 34, No. 12, pp. 1865-1875, 1962.

[4] Glass, J.R., "Nasal Consonants and Nasalized Vowels: An Acoustic Study and Recognition Experiment", S.M. Thesis, Massachusetts Institute of Technology, February 1985.

[5] Malécot, A., "Vowel Nasality as a Distinctive Feature in American English, *Language*, Vol. 36, No. 2, pp. 222-229, 1960.

[6] Rabiner, L.R., Schafer, R.W., *Digital Processing of Speech Signals*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1978.

[7] Randolph, M.A., "The Application of a Hierarchical Classification Technique to Speech Analysis." Paper presented at the 110th meeting of the Acoustical Society of America, Texas, 1985.

[8] Raphael, L., Dormann, M., Freeman, F., "Vowel and Nasal Duration as Cues to Voicing in Word-Final Stop Consonants: Spectrographic and Perceptual Studies, *Journal of Speech and Hearing Research*, Vol. 18, pp. 839-400, 1975.

[9] Tobias, J.V., "Relative Occurence of Phonemes in American English", *Journal of the Acoustical Society of America*, Vol. 31, p. 639, 1959.

Paper presented June 24, 1985, at IASTED Robotics & Automation Conference, Lugano, Switzerla

Automatic Alignment of Phonetic Transcriptions with Continuous Speech

Hong C. Leung and Victor W. Zue
Department of Electrical Engineering and Computer Science and
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139

ABSTRACT

This paper is concerned with the design and implementation of a system for automatic alignment of phonetic transcriptions with continuous speech. The implemented system consists of three modules. The speech signal is first segmented into broad classes using a non-parametric pattern classifier. Path finding techniques are then used to align the broad classes with the phonetic transcriptions. These aligned broad classes provide "islands of reliability" for more detailed segmentation and refinement of boundaries. Specific speech knowledge is utilized throughout the system. By doing alignment at the phonetic level, the system can often tolerate inter- and intra-speaker variability. The system was evaluated on seven hundred sentences, spoken by male and female speakers. 97% of the segments are mapped into only one phonetic event, approximately 80% of the time the offset between the boundary found by the automatic alignment system and a trained transcriber is less than 10 ms. Supporting software has also been developed so that final manual adjustments, if needed, can be made.

INTRODUCTION

Problem Statement

This paper describes a system that automatically aligns a phonetic transcription with the associated speech waveform. In developing such a system, we assumed that the speech signal has an underlying representation that consists of a sequence of phonetic segments. We recognize that during speech production, the acoustic realization of these phonetic segments may blend from one segment to another, due to the interaction among the various articulatory structures and their different degrees of sluggishness. The task of the system is then to associate a phonetic label with regions delineated by significant *acoustic landmarks*. We do not in any way imply that the alignment system finds the boundaries between phonetic segments.

The reliability of the acoustic landmarks in continuous speech is not at all uniform. Some landmarks are obvious and clear while others are more subtle. Figure 1 illustrates the spectrogram and various displays for the phrase, "Glue the sheet to the dark..", spoken by a male speaker. Row (a) of the Figure shows the phonetic transcription which is manually aligned by an experienced acoustic phonetician. As can be seen from the spectrogram, the transition from a strong fricative to a vowel, as in the word "sheet", is strongly evidenced by the abrupt decrease of high frequency energy and a sharp onset of low frequency energy. This kind of acoustic landmark is relatively easy to detect. On the other hand, the transition between a vowel and a sonorant as in the word "dark", is marked by more gradual acoustic changes. This second acoustic landmark is often quite subtle and is, in general, difficult to locate without first establishing the phonetic context. Therefore, the difficulty of the time alignment task will vary from one type of transition to another.

Motivation

Phonetic alignment is essential to many areas of speech research, since the time-aligned transcription can serve as pointers to specific phonetic events in the waveform. If a sufficient amount of time-aligned acoustic data is available, speech researchers will then be able to quantify the properties of phonetic segments and describe how their characteristics are modified by contexts. These results in turn will lead to a better model for speech production, as well as better rules for speech synthesis and recognition. A large database of aligned speech material is particularly important for phonetic recognition, since it can be used for phonetic knowledge acquisition, rule development, as well as system training and evaluation.

The automatic phonetic alignment system can also serve as a testbed for the development of specific phonetic recognition algorithms. It is well known that detailed phonetic recognition is extremely difficult, due to the context dependency of the acoustic realizations. In the automatic alignment task, we attempt to locate specific phonetic events when the identity and the contexts are known. Thus it can be viewed as a learning step towards phonetic recognition.

Traditionally, the alignment is done manually by a trained acoustic phonetician, who listens to the speech signal and visually examines various displays of the signal. There are several disadvantages to this approach. First, the task is extremely time consuming; even under the best of circumstances, the process of time alignment can take several minutes for one second of speech material. Second, the task requires the skill and knowledge possessed by a small number of experts. These two reasons combine to severely limit the amount of data that can be collected in this manner. Third, manual labeling often involves decisions that are highly subjective. Therefore, there is the lack of consistency and reproducibility of the results. Even if the sentence and the transcription were the same, the inter- and intra-transcriber variability can still be quite high. Finally, there is the problem of human error associated with tedious tasks.

Review of Literature

Over the past few years, several automatic time alignment procedures have been suggested in the literature. Most of these approaches attempt to align the speech waveform with a reference waveform, using dynamic programming algorithms. The reference waveform may be a known and previously labeled utterance [1], [3], a concatenation of stored templates [8], or a synthetically generated utterance [6]. In order for these methods to be effective, the two waveforms must not differ significantly in detailed phonetic structures, or the synthesis rules must be fairly advanced. A second approach, which also uses dynamic programming, is to segment and label the waveform into broad phonetic classes prior to time alignment [11]. A more detailed frame-by-frame labeling is then achieved by a second dynamic programming algorithm, using derivatives of energy and formant functions.

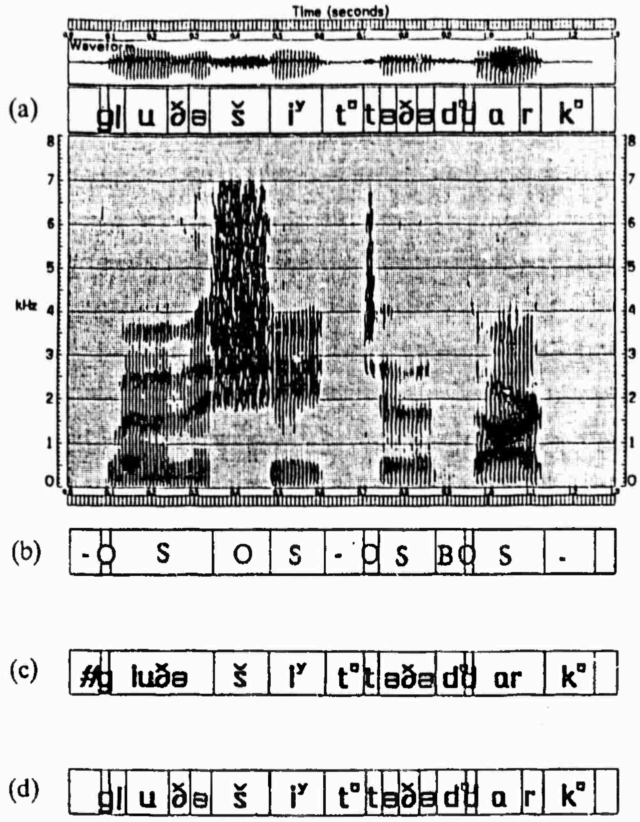


Figure 1: An example for the phrase, "Glue the sheet to the dark.."

This paper describes a new method of automatic phonetic alignment. This method utilizes a standard pattern classification algorithm, a path finding algorithm, and the constraints imposed by our acoustic-phonetic knowledge. The speech signal is first segmented into broad phonetic classes using a non-parametric pattern classifier. The resulting string is then aligned with the transcription using branch and bound search. Acoustic-phonetic knowledge is utilized extensively in the feature extraction for pattern classification, the specification of constraints for time-aligned paths, and the subsequent segmentation/labeling and refinement of boundaries.

SYSTEM DESCRIPTION

The basic structure of the system that we have developed is shown in Figure 2. The speech signal is digitized at 16 kHz and captured by an automatic end-point detection algorithm [5]. From the speech signal, a number of parameters are computed. These parameters are then used in conjunction with a pattern classifier to produce 5 broad phonetic classes. The output of the classifier is used to time-align major and robust acoustic events with the phonetic transcription. This initial time alignment serves as anchor points for subsequent detailed phonetic alignment utilizing a set of heuristic rules [7].

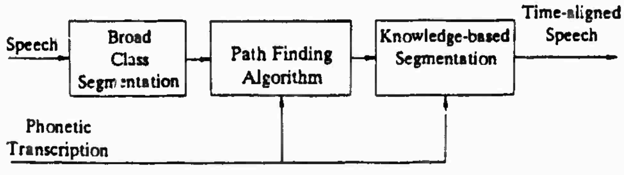


Figure 2: Basic system structure

Initial Broad Classification

In our opinion, direct alignment of the speech waveform with the detailed phonetic transcription is difficult, due to the high degree of acoustic variability in the speech signal. Our approach is to make an initial broad classification, relying on statistical pattern classification techniques. The objective is to determine robust acoustic-phonetic events that are relatively context-independent, and to use these events as anchor points for more detailed analysis. We have chosen to structure the classifier as a sequence of binary classifiers arranged in a binary decision tree. One advantage of using a set of classifiers is that a different feature vector can be used for each classifier in order to maximize the contrasts between the two possible output classes. For example, zero-crossing rate is helpful for distinguishing sonorants and obstruents, but not for distinguishing vowels from other voiced consonants. Thus the problem of classifying the speech signal into different classes can be reduced to a sequence of sub-problems, which are relatively easier to tackle.

At each node in the decision tree, a binary decision is made by a pattern classification machine as shown in Figure 3. The structure of each of the classifiers is identical; the only difference is the feature vectors and initial seed points used in the clustering algorithm. Each classifier starts with a set of M parameters selected based on acoustic-phonetic knowledge. (The number of parameters used, M, may be different for each of the binary classifiers.) The parameters are then processed through a seven-point median smoother, clipped, and normalized. Clipping is intended to emphasize the portions of the speech signal where boundaries are likely to occur. Clipping thresholds are determined statistically from a set of training data, such that segment boundaries fall within the transitional regions. Normalization then transforms each of the clipped feature parameters to the same scale. Together the clipping and normalization procedures effectively assign different weights to different feature parameters depending on how much the feature parameter distributions of the two classes overlap.

Every 5 ms, an M-dimensional feature vector is obtained. All the feature vectors for a given utterance constitute samples in the feature space. A binary decision is made in the M-dimensional feature space by using K-Means clustering, with a Euclidean distance metric [10]. It is well known that the location of the cluster centroids and the speed of convergence for a clustering algorithm depends on the choice of the initial seed points, the number of clusters, and the geometrical distributions of the data. The use of a binary classifier has the advantage that the algorithm always converges. In addition, the binary classifier enables us to apply our acoustic-phonetic knowledge and select initial seed points at the appropriate extrema of the feature space to maximize the contrast. We found that the algorithm typically converges after less than 10 iterations.

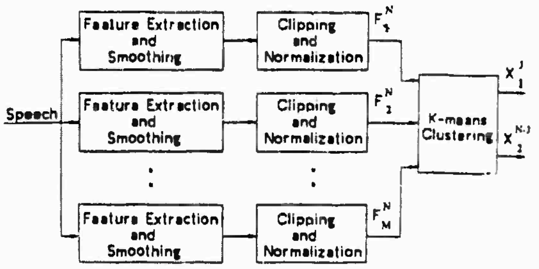


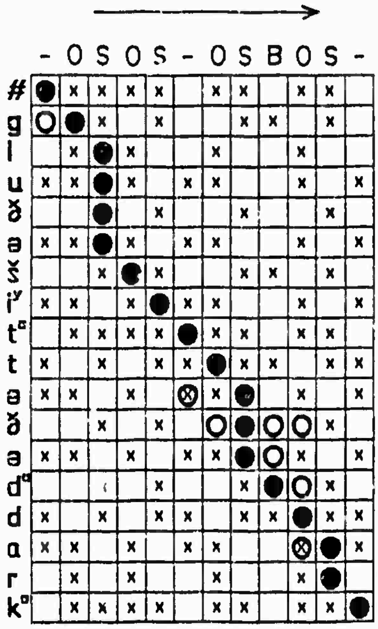
Figure 3: Block diagram of the pattern classification machine. Superscripts denote number of samples.

At the top of the decision tree, the clustering algorithm assigns one of two labels to every frame of data. Each class of data will pass through a different classifier at a lower node, and the process repeats. Our experience has shown that the broad phonetic classifier performs very well if the total number of classes is small, say 5 or 6. The performance of the classifier degrades substantially when one attempts to use it to make fine phonetic distinctions. In our implementation, the classifier assigns one of five labels to every frame of the data: S (vowel-like sonorant), O (obstruent), - (silence), B (nasals and voice-bars), and D (voiced consonants). Although rarely needed, a context-dependent median smoother is provided to remove spurious segments.

For the example shown in Figure 1, the output of the broad phonetic classifier is shown in row (b). Comparing the results with the spectrogram, we see that important and robust acoustic regions in this example have been found by the classifier.

Alignment

The output of the initial classifier is a broad, but presumably robust, description of the significant acoustic-phonetic events in the speech signal. In order to use this broad phonetic description as anchor points for more detailed analyses, the broad representation must now be aligned with the phonetic transcription. This is essentially a path finding problem, and we have chosen to use branch and bound search, where the path is heavily constrained by acoustic-phonetic rules [12]. Figure 4 illustrates how this is done for the same utterance as shown previously. The horizontal dimension represents the output of the classifier, while the vertical dimension represents the actual transcription. Durational information is used by the algorithm, but is not explicitly represented in this figure. Two kinds of constraints direct the algorithm to search for the correct path. First, the path is not allowed to traverse through certain cells, since this will produce implausible phonetic alignments. These mismatches are stored as a set of context-independent rules, and the resulting cells are marked in the figure by an x. For example, the first phoneme /t/ is not allowed to match a silence or a sonorant segment. Second, there is a set of rules that eliminates certain matches based on contextual information. The cells eliminated by the context-dependent rules are represented in the figure by the unfilled rectangles. For example, the first /t/ is not allowed to match the second obstrucnt due to a durational constraint. The remaining permissible cells are marked in the figure with filled or unfilled circles. The filled circles denote the optimum path, subject to a set of cost functions obtained through training. As can been seen from this example, the acoustic-phonetic constraints can often reduce the number of permissible paths dramatically.



	-	O	S	O	S	-	O	S	B	O	S	-
#	●	x	x	x	x		x	x		x	x	
g	○	●	x		x			x	x		x	
l		x	●	x			x			x		
u	x	x	●	x		x	x			x		x
ð			●		x			x			x	
ə	x	x	●	x		x	x			x		x
š			x	●	x			x	x		x	
iʸ	x	x		x	●	x	x			x		x
tᵒ		x	x	x	x	●	x	x		x	x	
t	x		x		x	x	●	x	x		x	x
ə	x	x		x		⊗	x	●		x		x
ð			x		x		○	●	○	○	x	
ə	x	x		x		x	x	●	○	x		x
dᵒ					x			x	●	○	x	
d	x		x		x	x		x	x	●	x	x
a	x	x		x		x	x			⊗	●	x
r		x		x			x			x	●	
kᵒ		x	x	x	x		x	x		x	x	●

Figure 4: Alignment mechanism for the same phrase, "Glue the sheet to the dark.."

Row (c) of Figure 1 shows the results of the time alignment obtained from the complete path shown in Figure 4. It can be seen that in some cases, there is only one phonetic event in a segment. In other cases, there can be as many as four phonetic events. Furthermore, a comparison with the hand transcription shows that all the phonetic events for this example are mapped to the correct broad class segment.

Knowledge-based Segmentation

The knowledge-based path finding algorithm divides the speech waveform into a sequence of segments. Each segment may be mapped to one or more phonetic events. No further processing is necessary if the matching is one or more segments to one phonetic event. For those segments which correspond to 2 or more phonetic events, more detailed segmentation is needed. This is done in two separate steps. Certain transitions between phonetic events, such as the transition between vowels and post-vocalic /r/'s, are not marked by reliable acoustic cues. In these cases, we have chosen to mark the boundary by a set of arbitrary but consistent rules. On the other hand, the transitions between some other phonetic events are more distinct. In these cases, further segmentation is achieved by the proper selection of feature parameters and algorithms based on contextual information. An example for this kind of phonetic segment is the intervocalic /ð/.

Row (d) of Figure 1 shows the results of the knowledge-based segmentation. We see that all the phonetic events in this example have been correctly located.

Application of Speech Knowledge

Throughout the development of the system, it was found that existing algorithms can be made more powerful by judicious application of speech knowledge. By structuring the initial classifiers in a binary decision tree, for example, specific acoustic attributes can be selected to enhance a particular phonetic contrast. Figure 5 shows a typical comparison of this procedure with an LPC-based classifier using the Itakura's distance metric [4]. In this figure, the output of the 4-way classifier has been converted to a level-coded waveform to facilitate visual comparison. The "feature-based" classifier consistently out-performs the "LPC-based" classifier in that the resulting classes are both stable and phonetically meaningful. We conclude that proper selection of acoustic measurements for classification has its payoffs.

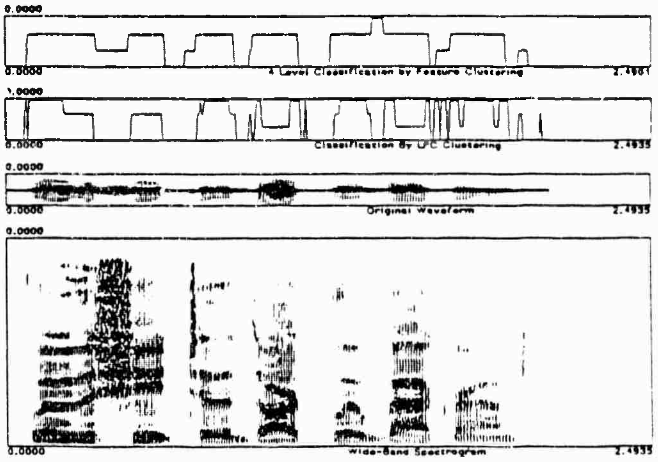


Figure 5: Performance comparison between clusterings based on LPC coefficients and pre-selected features

The path finding algorithm used in this system is not unlike the popular dynamic programming algorithm [9]. In this case, however, the alignment is performed at the segmental level, so that phonetic rules can be used to severely constrain the search space. In the example of Figure 4, the path alternatives, shown as open circles, are so limited that finding the correct alignment is often a trivial operation.

EVALUATION

Evaluation 1

The system was first evaluated using 40 distinct sentences, randomly chosen from the Harvard List of phonetically balanced sentences. Five talkers, three male and two female, each read twenty sentences, resulting in a total of one hundred (100) sentences. The corpus contains approximately 4 minutes of speech material and three thousand (3000) phonetic events. All sentences were hand transcribed by an experienced acoustic phonetician. For comparison, five of the one hundred sentences, selected at random, were manually labeled by a second transcriber. The entire process of manual labeling took upwards of 25 hours.

Figure 6 shows the percentage of phonetic events located after two separate stages of processing. A phonetic event is located when there is a correspondence between it and one or more acoustic segments. We see that approximately 80% of the phonetic events have been located after the alignment procedure. This number increased to 97% after knowledge-based segmentation.

Number of phonetic events in 1 segment	Path Finding Algorithm	Knowledge-based Segmentation
1	81%	97%
2	16%	2%
3 or more	3%	1%

Figure 6: Statistics on number of phonetic events in one segment after two different stages of processing.

We have also compared the reliability of the boundaries found by the system with those found by an experienced transcriber. This is done by computing the absolute difference between the two sets of boundaries. Figure 7(a) shows the cumulative distribution of the boundary offsets between the automatic alignment system and the acoustic phonetician. We see that approximately 75% of the boundaries are within 10 msec of each other, and over 90% of the boundaries are within 20 msec. Figure 7(b) shows the boundary offsets between the two transcribers for five of the sentences. Since it is difficult to assert exactly where a boundary should be, this curve provides a subjective indication of the performance of the alignment system. We see that the system-transcriber differences are similar in magnitude to the inter-transcriber differences.

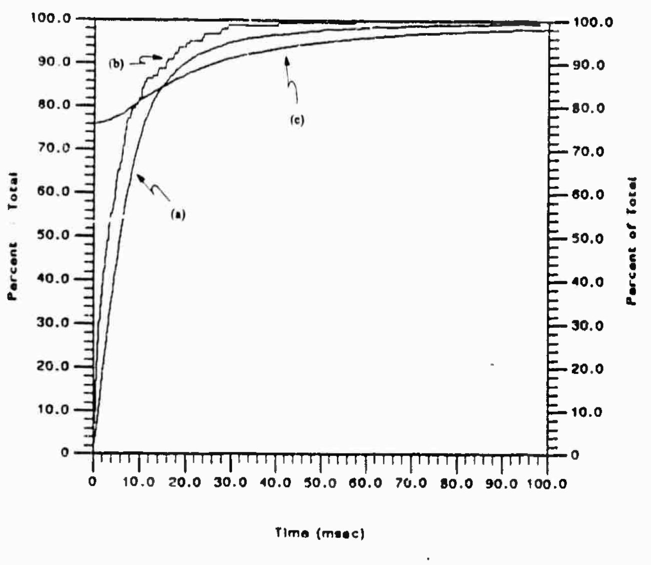


Figure 7: Cumulative distributions of the boundary offset

Evaluation 2

The system was also evaluated on 300 distinct sentences, randomly chosen from the Harvard List. Sixty speakers, thirty male and thirty female, each read ten of the sentences, resulting in a total of 600 sentences, and approximately 25 minutes of speech material.

Instead of asking a human transcriber to label these sentences for system evaluation, the output of the alignment system was checked by the same acoustic phonetician as in Evaluation 1. Misalignments, when present, were then corrected by hand. Figure 7 (c) shows the boundary offsets between the sets of boundaries before and after hand correction. We can see that 76% of the boundaries do not need to be corrected, whereas over 80% of the boundaries are within 10 msec. In other words, the performance results for the two databases were quite similar.

SUMMARY

In this paper we described a system that automatically aligns a phonetic transcription with the corresponding speech waveform. The system performs initial classification by a pattern classification algorithm. The output of the classifier is used to determine "islands of reliability" for further segmentation. By proper application of speech knowledge, the performance of the system can be improved. The entire system runs in approximately 35 times real time on our lisp machine workstations. In addition, supporting software for entering transcriptions and correcting alignment is also provided. We have now used the system for the collection of over 1000 sentences. Some of the output of the alignment system has already been used for different research projects [2]. We are encouraged by the results, and are hopeful that this system will play a major role in establishing a large database for speech research.

[This research was supported by the Office of Naval Research under contract N00014-82-K-0727.]

Bibliography

[1] Chamberlain, R.M., Bridle J.S., "ZIP: A Dynamic Programming Algorithm for Time-aligning Two Indefinitely Long Utterances", *ICASSP 1983*.

[2] Glass J.R., Zue V.W., "Analysis and Recognition of Nasal Consonants in American English", Seventh IASTED International Symposium, *Robotics and Automation 1985*, Lugano, Switzerland.

[3] Hohne et al., "On Temporal Alignment of Sentences of Natural and Synthetic Speech", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol.ASSP-31, No. 4, August 1983.

[4] Itakura F., "Minimum Prediction Residual Applied to Speech Recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing* Vol.ASSP-23, February 1975.

[5] Lamel L.F. et al., "An Improved Endpoint Detector for Isolated Word Recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol.ASSP-29, No. 4, August 1981.

[6] Lennig M., "Automatic Alignment of Natural Speech with a Corresponding Transcription", *Speech-Communication, 11th International Congress on Acoustics, Toulouse*, July 1983.

[7] Leung H.C., "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech", *S.M. Thesis, Massachusetts Institute of Technology* 1985.

[8] Lowry M. R., "Automatic Labelling of Speech from the Phonetic Transcription", *S.M. Thesis, Massachusetts Institute of Technology* 1978.

[9] Rabiner L.R., Myers C.S., "Connected Digit Recognition Using a Level-Building DTW Algorithm", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol.ASSP-29, No. 3, June 1981.

[10] Tou J.T., Gonzalez R.C., *Pattern Recognition Principles*, Addison-Wesley 1974.

[11] Wagner, M., "Automatic Labelling of Continuous Speech with a Given Phonetic Transcription Using Dynamic Programming Algorithms", *ICASSP 1981*.

[12] Winston, P.H., *Artificial Intelligence*, Addison-Wesley 1984.

DISTRIBUTION LIST

	DODAAD Code	
Head Information Sciences Division Office of Naval Research 800 North Quincy Street Arlington, Virginia 22217	N00014	(1)
Administrative Contracting Officer E19-628 Massachusetts Institute of Technology Cambridge, Massachusetts 02139		(1)
Director Naval Research Laboratory Washington, D.C. 20375 Attn: Code 2627	N00173	(1)
Defense Technical Information Center Bldg. 5, Cameron Station Alexandria, Virginia 22314	S47031	(12)